Infrared Invisible Clothing: Hiding from Infrared Detectors at Multiple Angles in Real World

Xiaopei Zhu[1,2] Zhanhao Hu[2] Siyuan Huang[2] Jianmin Li[2] Xiaolin Hu[2,3,4*]
[1]School of Integrated Circuits, Tsinghua University, Beijing, China
[2]Department of Computer Science and Technology, Institute for Artificial Intelligence,
State Key Laboratory of Intelligent Technology and Systems, BNRist, Tsinghua University, Beijing, China
[3]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China
[4]Chinese Institute for Brain Research (CIBR), Beijing, China
{zxp18, huzhanha17}@mails.tsinghua.edu.cn
{siyuanhuang, lijianmin, xlhu}@mail.tsinghua.edu.cn

Abstract

Thermal infrared imaging is widely used in body temperature measurement, security monitoring, and so on, but its safety research attracted attention only in recent years. We proposed the infrared adversarial clothing, which could fool infrared pedestrian detectors at different angles. We simulated the process from cloth to clothing in the digital world and then designed the adversarial "QR code" pattern. The core of our method is to design a basic pattern that can be expanded periodically, and make the pattern after random cropping and deformation still have an adversarial effect, then we can process the flat cloth with an adversarial pattern into any 3D clothes. The results showed that the optimized "QR code" pattern lowered the Average Precision (AP) of YOLOv3 by 87.7%, while the random "QR code" pattern and blank pattern lowered the AP of YOLOv3 by 57.9% and 30.1%, respectively, in the digital world. We then manufactured an adversarial shirt with a new material: aerogel. Physical-world experiments showed that the adversarial "QR code" pattern clothing lowered the AP of YOLOv3 by 64.6%, while the random "QR code" pattern clothing and fully heat-insulated clothing lowered the AP of YOLOv3 by 28.3% and 22.8%, respectively. We used the model ensemble technique to improve the attack transferability to unseen models.

1. Introduction

Thermal infrared ("infrared" for short throughout the paper) imaging is widely used in many areas such as human temperature measurement, safety monitoring, and autopilot. Infrared imaging has its unique advantages [42]. First of all, infrared imaging can image in the dark, which means that the infrared equipment can work 24 hours a day. Secondly, unlike radar imaging, infrared imaging does not need to transmit signals actively, which saves more energy. Third, infrared imaging can measure temperature information, which is unavailable in visible light or radar filed. During the COVID-19 pandemic, infrared imaging is widely used for fast body temperature monitoring.

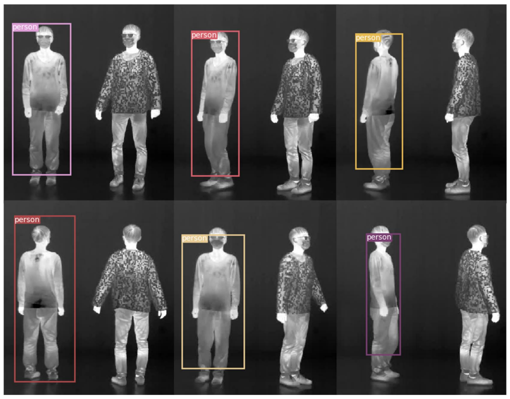


Figure 1. Demonstration of physical infrared attack. A person wearing the adversarial clothes hid from infrared detectors at multiple angles. Whereas, the other person wearing ordinary clothes was detected (indicated by bounding boxes).

Infrared object detection combines deep learning with infrared imaging technique. The security of deep learning has attracted more and more attention in recent years. Szegedy et al. [33] found that neural networks can output error results with high confidence by adding specially crafted perturbations to the input data. The perturbed data is

[*]Corresponding author.

called adversarial example. Adversarial examples threaten not only the digital world [4, 13, 21, 22, 33, 38] but also the physical world [1, 10, 32, 34, 39]. Nowadays, most research on adversarial examples focuses on the visible light field; some are in the radar and infrared field. This paper focuses on the security of thermal infrared object detection systems.

Traditional infrared stealth generally uses two methods: heat insulation and active cooling, but they usually cannot completely hide thermal infrared signals. For example, as people need to breathe, the human body always emits thermal infrared signals to the outside. Adversarial example technology provides a different way of stealth, which can make deep learning-based detectors unable to detect people through carefully designed patterns [18, 34, 39]. Visible light patterns can be easily displayed in the physical world through printing or LED displays, but infrared patterns are difficult to be "printed".

Zhu et al. [42] proposed a physical method using small bulbs to attack infrared pedestrian detectors. To the best of our knowledge, that was the first work to realize physical attacks on the thermal infrared pedestrian detectors. But that method has an obvious shortcoming. The small bulb board can only attack at a specific angle (usually the front) of the human body. In this work, our goal is to solve this problem by designing a new physical attack method. Specifically, we want to design a piece of clothing to achieve a "wearable" attack. There are two requirements for the adversarial clothing. First, the designed clothes should have a specific texture and deceive infrared pedestrian detectors from different angles. Second, the adversarial pattern on the clothing should still remain effective after non-rigid deformation.

Towards this goal, we designed infrared adversarial clothes based on a new material: *aerogel*. The core of our method is to design a basic pattern that can be expanded periodically, and make the pattern after random cropping and deformation still have an adversarial effect, so that we can process the flat cloth with adversarial pattern into any 3D clothes. Figure 1 shows an example of physical infrared clothing attack and control experiments. The contributions of this paper are as follows:

- First, we simulated the process from cloth to clothing in the digital world and then designed the adversarial "QR code" pattern.
- Second, we manufactured infrared adversarial clothing based on a new material aerogel, which hid from infrared detectors at multiple angles in the physical world.

2. Related Works

2.1. Digital Adversarial Attacks

Since Szegedy et al. [33] discovered the vulnerability of deep neural networks, many digital attack methods have been proposed. Classic digital attack methods include gradient-based methods (e.g., FGSM [14], BIM [21], DeepFool [26], PGD [25]), optimization-based methods (e.g., L-BFGS [33], C&W [4], ZOO [5]), and GAN-based methods (e.g., AdvGAN [38], PS-GAN [22], AdvFaces [8]). The digital attack methods assume that the attacker can modify the model's input, which is difficult to achieve in the real-world setting. Some recent works [6, 41] used 3D modeling to simulate real-world attacks, but there was still a gap between 3D modeling and real scenes.

2.2. Physical Adversarial Attacks

Most physical attacks focused on the visible light and the rader field. These attacks can be roughly divided into classification attacks and detection attacks. For classification attacks, Athalye et al. [1] successfully deceived the classification model with a 3D printed tortoise. Eykholt et al. [10] proposed Robust Physical Perturbations (RP2). Duan et al. [9] proposed Adversarial Laser Beam, which could quickly attack the classification model in the physical world in a non-contact manner.

For detection attacks, Thys et al. [34] printed the adversarial pattern on a piece of paper and successfully made the YOLOv2 [27] unable to detect people. Xu et al. [39] designed a T-shirt with an adversarial pattern printed on the front. Huang et al. [18] proposed Universal Physical Camouflage (UPC) to fool Faster-RCNN [30] in the physical world. Hu et al. [17] used the generative adversarial network (GAN) to generate natural looking adversarial patches while maintaining high attack performance. Tu et al. [35] proposed a method to generate 3D adversarial mesh to fool LiDAR detectors. A recent study [3] shows that the designed adversarial 3D-printed object could be invisible for both camera and LiDAR.

To the best of our knowledge, only one work focused on the safety of infrared object detection. Zhu et al. [42] designed a board decorated with small bulbs to attack infrared pedestrian detectors. The person holding the adversarial board could be invisible to the infrared detection model.

2.3. Infrared Stealth Materials

Infrared stealth materials can be roughly divided into low emissivity materials and temperature control materials. Aluminum is a commonly used material with low emissivity, but it is easily oxidized. Fan et al. [11] synthesized a new Al-reduced graphene oxide composite material, which had improved anti-oxidability and had excellent infrared stealth capabilities. Temperature-controlled infrared stealth material realizes infrared stealth by reducing the surface temperature. Shang et al. [31] studied the microstructure and thermal insulation property of silica composite aerogel, which showed good thermal insulation stability at room temperature. Wang et al. [36] proposed a

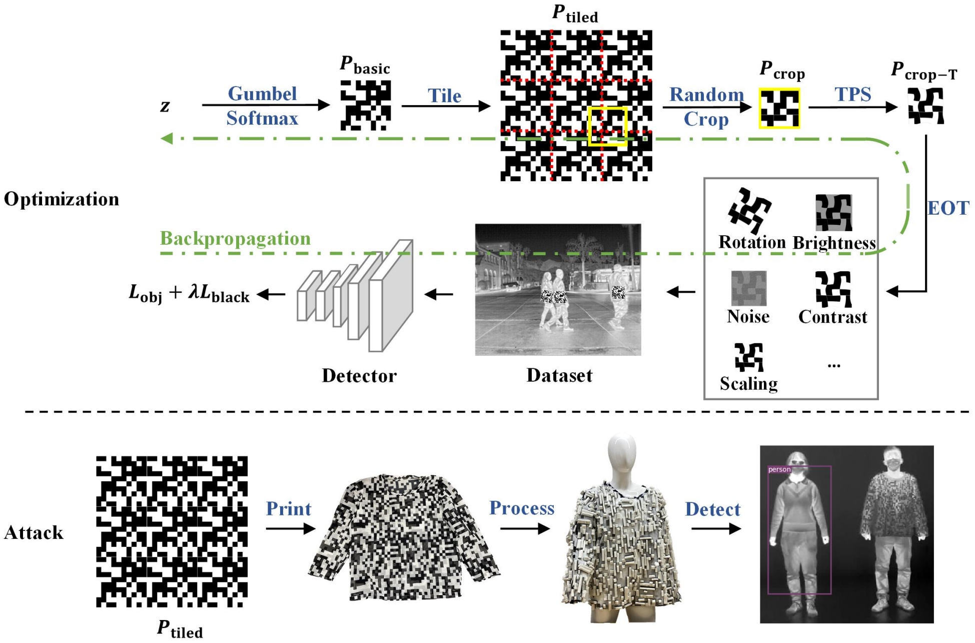


Figure 2. Pipeline of proposed method. Top: attack in the digital world by optimizing a binary pattern. Bottom: attack in the physical world.

polysiloxane bonded silica aerogel with enhanced thermal insulation capability.

3. Methods

3.1. Simulating the Cloth-to-Clothing Process in the Digital World

Our goal is to make a piece of clothing with adversarial texture in infrared imaging. It is required that this piece of clothing will have a certain adversarial effect from any angle. First of all, let's review the real-world process of making clothes. In the real world, we first look for a piece of cloth. We usually design a basic pattern and periodically expand the basic pattern on the plane. After we get the cloth printed with the expanded pattern, we crop and tailor it to clothing.

We simulate the process from the cloth to clothing in the digital world, as shown in Figure 2. Let P_{basic} denote the basic pattern unit and P_{tiled} denote the image after tiling (periodic expansion) of P_{basic}. P_{tiled} can be any size. We define the tiling function as TILE. The process above can be expressed as

$$P_{\text{titled}} = \text{TILE}\left(P_{basic}\right). \tag{1}$$

When we take photos of people wearing clothes, we always photograph a part of the clothing, which can be regarded as cropping from the entire original cloth from which the clothing is made. We define a function RC for random cropping. The randomness here has two meanings: the randomness of the crop position and randomness of the crop size. Let P_{crop} denote the patch randomly cropped from the pattern P_{tiled}, namely

$$P_{crop} = \text{RC}\left(P_{tiled}\right). \tag{2}$$

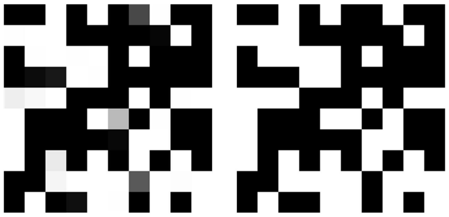
Due to the irregular deformation of cloth in the real world, we use the Thin Plate Spline (TPS) [37] interpolation method to approximate this process. TPS is an algorithm to simulate planar non-rigid deformation, especially suitable for cloth. Its basic idea is to give K matching points in two images and make the points of one image with specific deformation correspond exactly to the points of the other image. Let P_{crop-T} denote the patch after TPS transformation, namely

$$P_{crop-T} = \text{TPS}(P_{crop}). \tag{3}$$

To simulate the disturbance in the real-world environment, like lighting changes, we used the Expectation over Transformation (EOT) [1] method.

$$P_{crop-TE} = \text{EOT}\left(P_{crop-T}\right). \tag{4}$$

EOT randomly changes the position, brightness, contrast, rotation angle, scale of the patch P_{crop-T}, and adds random noise to simulate changes in the real world as realistically as possible.



(a) (b)

Figure 3. The comparison between (a) the patch with approximate pixel values and (b) the patch with the real pixel values.

3.2. Design of Binary Pattern

We then consider the design of the cloth pattern. The mechanism of infrared imaging is quite different from that of visible light imaging. An infrared image is a grayscale image. The pixel value reflects the temperature of the object's surface. The higher the pixel value, the higher the temperature. We design the cloth pattern as shown in Figure 2, which looks like a Quick Response (QR) code. The white pixel reflects the normal body surface temperature, and the black pixel reflects the surface temperature after using thermal insulation materials. Therefore, we transform cloth pattern design into a binary optimization problem. For each pixel in the "QR code", we only need to consider whether to put heat insulation material here or not. This design is beneficial to subsequent physical implementation.

However, since each pixel in the "QR code" is a binary value, the "QR code" pattern cannot be directly optimized by the gradient descent method. To solve this problem, we use the Gumbel-softmax technique [19]. The details are as follows. For each pixel in the image, π_0 denotes the probability of being black, and the probability of being white is π_1. Clearly, $\pi_0 + \pi_1 = 1$. First, we introduce Gumbel noise. The purpose of adding Gumbel noise is to add randomness to the sampling operation. g_i denotes the Gumbel noise of this pixel.

$$g_i = -\log\left(-\log\left(u_i\right)\right), u_i \sim Uniform\left(0,1\right). \qquad (5)$$

We next calculate the vector y_i $(i = 0, 1)$ used for sampling. $[y_0, y_1]^\top$ is an approximate representation of the one-hot vector. τ is a hyperparameter. $[y_0, y_1]^\top$ is closer to the one-hot vector when τ is smaller. We choose $\tau = 0.1$ in our experiment.

$$\begin{aligned} y_i &= \mathrm{Softmax}\left[\left(g_i + \log \pi_i\right)/\tau\right] \\ &= \frac{\exp\left(\left(g_i + \log \pi_i\right)/\tau\right)}{\sum\limits_{i=0}^{1} \exp\left(\left(g_j + \log \pi_j\right)/\tau\right)}. \end{aligned} \qquad (6)$$

Next, we assume that the pixel value of the grayscale image is in the interval $[0, 1]$, and we can find the approximate value $\tilde{p}$ at that pixel, $\tilde{p} = y_0 \times 0 + y_1 \times 1 = y_1$. Since $\tilde{p}$ is differentiable relative to π_i, we can optimize $\tilde{p}$ using the gradient descent method. The corresponding relationship between the real value p and the approximate value $\tilde{p}$ at this pixel is:

$$p = \begin{cases} 1, \tilde{p} \geq 0.5 \\ 0, \tilde{p} < 0.5. \end{cases} \qquad (7)$$

In the optimization process, we use $\tilde{p}$ to approximate p due to the need for gradient information; in the attack process, we use p directly. Figure 3 is a comparison between the patch with approximate pixel values and the patch with the real pixel values. It shows that the Gumbel-softmax technique can effectively help the approximation of binary images.

3.3. Optimizing the Binary Pattern

Figure 2 shows the optimization process. P_{basic} is an $N \times N$ patch. The variable z in the hidden space is the set of probability values π_i $(i = 0, 1)$ of each pixel in the patch P_{basic}, and the size of z is $2 \times N \times N$. P_{basic} is tiled to P_{tiled}. P_{crop} is randomly cropped from P_{tiled}. P_{crop-T} is the patch after TPS transformation. $P_{crop-TE}$ is formed by P_{crop-T} through EOT method. $P_{crop-TE}$ is pasted on the pedestrians in the data set, and then we input the patched images into the object detector. We update the variable z according to the loss function and further update the patch P_{basic}.

Our loss function has two parts, L_{obj} and L_{black}:

$$L = L_{obj} + \lambda L_{black}. \qquad (8)$$

The parameter λ (> 0) is determined empirically. We explain L_{obj} and L_{black} in what follows.

L_{obj} denotes the object score of the object detector. Let x denote the original image in the dataset and $\tilde{x}$ denote the patched image. Let f denote a model, θ denote its parameters. $f\left(x, \theta\right)$ denotes the model's outputs given the input x. Most object detectors have three outputs: position of the bounding box $f_{pos}\left(x, \theta\right)$, object probability $f_{obj}\left(x, \theta\right)$, and class probability $f_{cls}\left(x, \theta\right)$. Our goal is to make object detectors unable to detect pedestrians, so we want to lower the $f_{obj}\left(x, \theta\right)$ as much as possible. To fool the object detectors in the real world, we consider various transformations of the patches during attack, including translation, rotation, scale, noise, contrast, and brightness. Furthermore, we try to achieve a universal attack on different pedestrians due to the intraclass variety of pedestrians. Let $\mathbb{T}$ denote the set of transformations, $\tilde{x}_t$ denotes the patched image considering patch transformations. The data set has M images. Considering all above factors, L_{obj} can be described as

$$L_{obj} = \frac{1}{M} \sum\nolimits_{i=1}^{M} \mathrm{E}_{t \in \mathbb{T}} f_{obj}^{(i)}\left(\tilde{x}_t, \theta\right). \qquad (9)$$

L_{black} is the average probability of black pixels appearing in patch P_{basic}. The reason to propose this loss function is as follows. In physical implementation, the black pixels correspond to the heat-insulating material. The fewer black pixels are, the less heat-insulating material we use. This not only saves material, but also improves the air permeability and comfort of the clothes. From Section 3.2, we know that the probability of a single-pixel being black is π_0. For an $N \times N$ patch P_{basic} , the average probability of black pixels is:

$$L_{black} = \frac{\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \pi_0 (i,j)}{N \times N}. \tag{10}$$

For the ensemble attack [24], we aim to lower each detector's objectness score at the same time. We assume there are F detectors, and the objectness score of i-th detector is $L_{obj}^{(i)}$. We take the sum of these losses. Thus, the total loss of the ensemble attack is

$$L_{ensemble} = \sum_{i=1}^{F} L_{obj}^{(i)} + \lambda L_{black}. \tag{11}$$

3.4. Physical Implementation

Infrared stealth materials can be roughly divided into low emissivity materials and temperature control materials. According to the Stefan-Boltzmann law [7], the infrared radiation is more sensitive to temperature, so we gave priority to temperature control materials. We tested two common fabrics (cotton and polyester), two thermal insulation tapes (Teflon and polyimide), and a new type of material (aerogel). See *Supplementary Material* for their photos. We aimed to find a material that has the best thermal insulation performance.

The process for making clothes is as follows. First, we printed the "QR code" pattern we designed on a $1.5m \times 1.5m$ cloth. Next, We hired a tailor to make the cloth into a piece of clothing. Then we cropped the infrared stealth material into blocks and stuck them on the black area of the clothes.

4. Experiments

4.1. Dataset

We used the $FLIR\_ADAS\_v1\_3$ dataset [12] released by FLIR company. $FLIR\_ADAS\_v1\_3$ provides an annotated thermal image set for training and validation of object detection. The original dataset contains four types of objects, namely people, dogs, cars and bicycles. Since we focused on people, we filtered the dataset and selected 9900 images that contained people. We named the subset $PEOPLE\_FLIR$. The training set contained 7873 images, and the test set contained 2027 images.

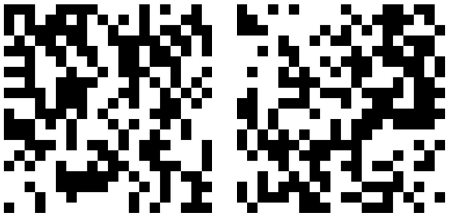


Figure 4. Optimized "QR code" texture based on (a) YOLOv3 (b) ensemble models.

4.2. Target Detector

We followed the work of Zhu et al. [42] and chose the same target detector YOLOv3 [28] . Kristo et al. [20] compared the performance of state-of-the-art infrared detectors such as Faster-RCNN [29], Cascade-RCNN [2], SSD [23], and YOLOv3. They found that YOLOv3 was significantly faster than other detectors while achieving performance comparable with the best. We resized the input images to 416×416 as required by YOLOv3. We used the pretrained weights officially provided by YOLO and then fine-tuned them on $PEOPLE\_FLIR$. The target model's AP was 97.27% on the training set and 85.01% on the test set. In our experiments, we first attacked YOLOv3 and then attacked other detectors under the black box setting.

4.3. Simulation of Physical Attacks

4.3.1 Attack YOLOv3 in the Digital World

As mentioned in Section 3.3, the size of variable z is $2 \times N \times N$. P_{basic} is an $N \times N$ patch. In our experiment, we chose $N = 20$ (See section 4.3.2 for the results of other values). We expanded the side length of P_{basic} by 5 times, so the size of P_{tiled} was 100×100. P_{crop} was randomly cropped from P_{tiled}, the crop size was randomly sampled from [10,30]. We took matching points $K = 16$ in TPS transformation. The set of EOT transformations included changes of patch position, brightness, contrast, rotation, angle, scale, and noise.

Next, we used the training set of $PEOPLE\_FLIR$, and placed $P_{crop-TE}$ in a random position of the human body according to the bounding box. The proportion of the patch size to the height of the bounding box varied from 0.1 to 0.3 according to the crop size. Next, we inputted these patched images into YOLOv3. We used a stochastic gradient optimizer with momentum. The optimizer used the backpropagation algorithm to update the parameters of variable z by minimizing Equation 8, and further updated the patch P_{basic}. The hyper-parameter λ in loss function was 0.1 (the sensitivity of this parameter is analyzed in Section 4.3.3). See *Supplementary Material* for details about the hyperparameter setting such as batch size, learning rate, etc. Figure

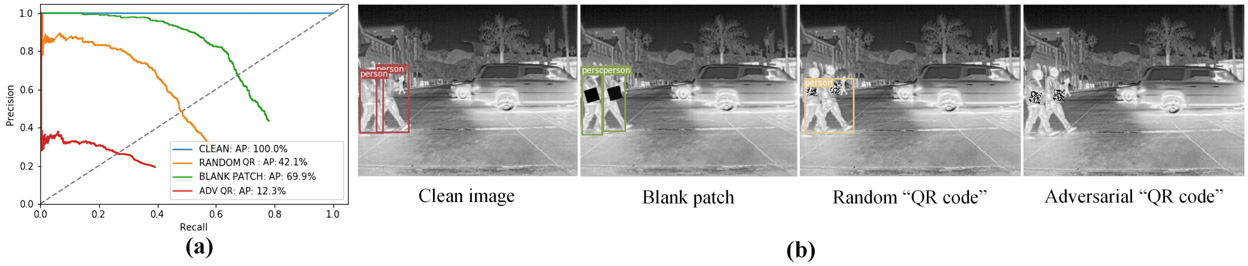


Figure 5. Digital attacks. (a) Evaluation of digital attacks. (b) Examples of digital attacks. Bounding boxes indicate successful detecting of persons.

4(a) shows the optimization result.

Next, we applied the optimized pattern (Figure 4(a)) to the test set in the same way as the optimization process. To further compare the effect of the attack, we used a random “QR code” pattern and a blank pattern for control experiments. In our experiment, the pixel value of the blank pattern was 0, which corresponded to the situation where the heat was completely insulated in the real world. We applied these patterns to the test set of $PEOPLE\_FLIR$, and then inputted the patched images to YOLOv3 to test their attack performance. We defined the model's output of clean images input as ground truth (GT). We used the Intersection over Union (IOU) method to calculate the detection accuracy. The precision-recall (P-R) curves are shown in Figure 5(a). The results showed that the optimized “QR code” pattern made the average precision (AP, the area under the PR curve) of YOLOv3 drop by 87.7%. In contrast, the random “QR code” pattern and blank pattern made the AP of YOLOv3 drop by 57.9% and 30.1%, respectively. Although random “QR code” pattern and blank pattern also lowered AP of the model, it is far less effective than the optimized pattern. Figure 5(b) shows some examples in the digital world.

4.3.2 Effect of Resolution of the Basic Patch

In the previous experiment (Section 4.3.1), the resolution of P_{basic} was 20×20. We then studied the resolution of 10×10, 20×20, 30×30, 40×40 and 50×50. The pattern optimization and testing process were described in Section 4.3.1. Table 1 showed the AP decrease of YOLOv3 corresponding to different resolutions. The more AP decrease meant the better attack effect. The result showed that the attack performance decreased if the resolution was too large or too small. And 20×20 was our best result.

Table 1. Effect of Resolution of Basic Patch

Patch size	10	20	30	40	50
AP decrease	64.9%	87.7%	85.6%	83.6%	83.4%

4.3.3 Effect of the Parameter λ in the Loss Function

The parameter λ in Equation 8 balances L_{obj} and L_{black}. We studied lambda values of 0, 0.01, 0.05, 0.1, 0.5, and 1. We adopted the same optimization and testing methods as described in Section 4.3.1. We evaluated the attack performance of patterns generated with different λ values by AP decrease. We also calculated the proportion of black pixels in different patterns. The results are shown in Table 2. The results showed that there was a certain trade-off between improving the attack performance and reducing the proportion of black pixels in the basic patch. The higher the proportion of black blocks, the stronger the attack performance of patterns under certain conditions.

4.4. Attacks in the Physical World

4.4.1 Physical Test of Thermal Insulation Materials

We tested the thermal insulation performance of the five materials stated in Section 3.4. See *Supplementary Material* for more details. The results showed that the aerogel had good thermal insulation properties and remained stable over time.

4.4.2 Attack YOLOv3 in the Physical World

We cropped the aerogel felt (Figure 6(a)) into many blocks and stuck them on the clothes. The manufacturing process is in the *Supplementary Video*. The manufacturing cost of our whole piece of clothing was within 50 USD, which meant the possibility of mass production. The finished clothing was shown in Figure 6(b). For better comparison, we also made a piece of random “QR code” pattern clothing (Figure 6(c)) and a piece of fully heat-insulated clothing (Figure 6(e)). These clothes had the same size. Next, we tested the attack performance of these clothes in the real world.

The infrared camera we used was FLIR T630sc (FPA 640 × 480, NETD<40mK). We invited 7 volunteers to participate in our experiment. This experiment was approved

Table 2. Effect of Parameter λ in the Loss Function

λ	0	0.01	0.05	0.1	0.5	1
Black pixel ratio	52.0%	49.8%	48.2%	47.3%	45.3%	44.3%
AP decrease	89.2%	88.4%	88.2%	87.7%	87.6%	87.2%

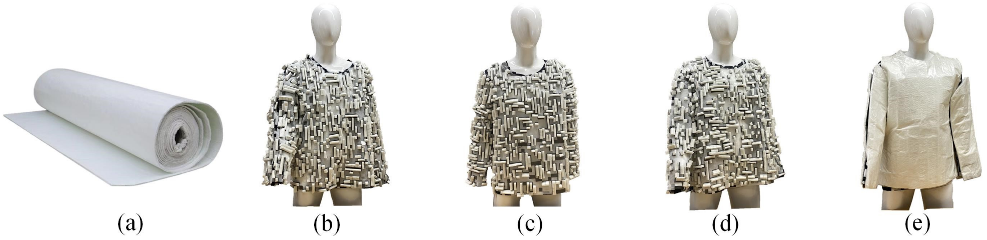


Figure 6. Physical material and clothing. (a) Aerogel felt. (b) Adversarial clothing based on YOLOv3. (c) Random "QR code" clothing. (d) Adversarial clothing based on ensemble models. (e) Fully heat-insulated clothing.

by the Institutional Review Board (IRB). The volunteers wore adversarial "QR code" clothing, random "QR code" clothing, fully heat-insulated clothing or ordinary clothing. We photographed volunteers in multiple scenes indoors and outdoors, and the camera's distance from them varied within 1-15 meters. At the same time, they can change different postures according to their preferences, such as standing, sitting, and even reclining, etc. We photographed the volunteers in different scenes simultaneously and sent these infrared images to YOLOv3. The threshold of the detection output was 0.7. Figure 7 gives some specific examples. We can see that people wearing adversarial "QR code" clothing were not detected by YOLOv3 even if they were in different postures, keeping different distances from the camera, and in different scenes. While at the same time, people wearing random "QR code" clothing, fully heat-insulated clothing or ordinary clothing were detected. The results showed the effectiveness of our method in the physical world. See *Supplementary Video* for the demo.

To quantitatively evaluate the effect of physical attacks, we recorded 120 videos in different scenes. 60 videos were recorded indoors, and the others were recorded outdoors. 5 volunteers were the actors in the video. We fixed the camera's position. Then we selected three typical positions which were 3 meters, 5 meters, and 7 meters from the camera to test attack performance. Then we invited volunteers to rotate in situ at a constant speed in these positions. For fair comparison, the same volunteer needed to wear adversarial "QR code" clothing, random "QR code" clothing, fully heat-insulated clothing and ordinary clothing, respectively, in the same position. We recorded videos with our infrared camera. We sampled the videos (from 7 volunteers) at 3 frames per second, and got 900 frames per condition, which made 3600 frames in total. We used manual annotation as the ground truth (GT) and then used the output of YOLOv3 to calculate AP by IOU method. The results showed that the adversarial "QR code" clothing reduced the detector's AP by 64.6%, while the random "QR code" clothing, fully heat-insulated clothing and ordinary clothing made the detector's AP drop by 28.3%, 22.8% and 4.0%, respectively.

We then analyzed the attack effect at different angles. We chose a distance of 5m from the camera, and marked some key angles on the ground. We then took 11 sample points from a counterclockwise rotation angle of 0° to 180° for statistics. We selected 100 frames at each angle, and used the attack success rate (ASR, the ratio of the number of frames that were not detected to the total number of frames) as the evaluation method. The threshold of the detector was 0.7. As shown in Figure 8(a), we plot the curve of the ASR with different angles. When the volunteers had their front to the camera (0°) or back to the camera (180°), the ASR was the highest. When they had their side to the camera (90°), the ASR was the lowest. One possible reason is that the area of the adversarial pattern on the side (90°) was relatively small.

We then studied the relationship between the ASR and distances. We kept the clothes always towards (0°) the camera, and the distance from the camera varied from 3 to 10 meters. We took 8 sample points between 3 and 10 meters for statistics, and selected 100 frames at each position. We plot the curve of the ASR with different distances (Figure 8(b)). When the distance was between 3 and 7 meters, the ASR was above 0.8. When the distance exceeded 7 meters, the ASR dropped faster, because the adversarial pattern was not easy to attack successfully in a smaller view.

4.5. Ensemble Attack

We found attack transferability of the single model was limited. The pattern optimized on YOLOv3 only lowered

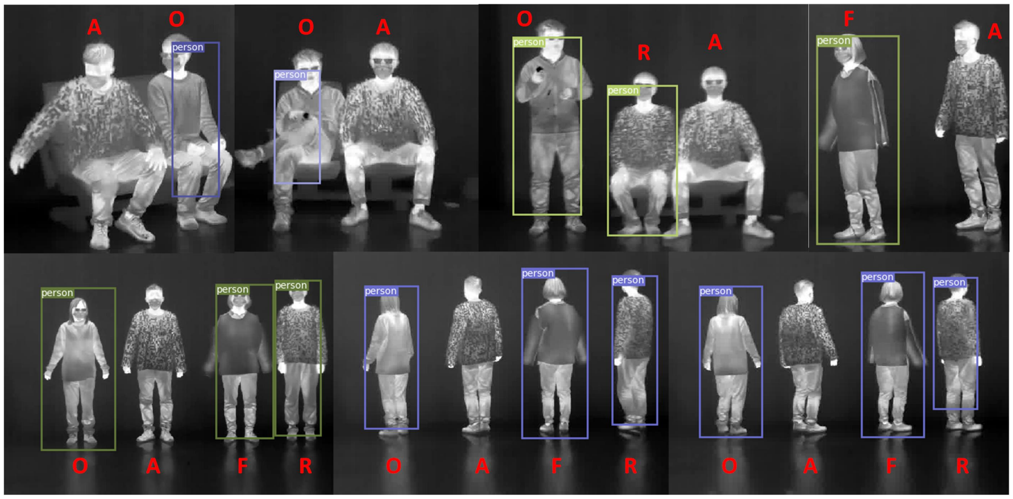
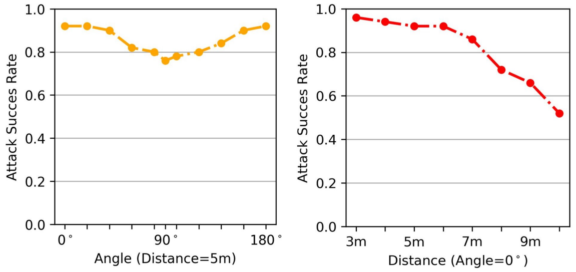


Figure 7. Visulization results of physical attacks. Persons wearing different clothing could be in various poses. A: adversarial "QR code" clothing. R: random "QR code" clothing. F: fully heat-insulated clothing. O: ordinary clothing.



(a) (b)

Figure 8. Analysis of attacks at different (a) angles and (b) distances

the AP of Deformable DETR, RetinaNet, and Libra-RCNN by 13.7%, 25.3%, and 32.7% in the digital world, respectively. We then used the model ensemble technique [24] as described in Section 3.3. Figure 4(b) shows a pattern obtained by ensembling YOLOv2, YOLOv3, Faster-RCNN, and Mask-RCNN [16] during the optimization process. It caused the AP of Deformable DETR, RetinaNet, and Libra-RCNN to drop by 24.7%, 58.2%, and 66.9% in the digital world. Then we manufactured a piece of clothing with the pattern obtained by model ensembling in the physical world (see Figure 6(d)), which made the AP of Deformable DETR, RetinaNet, and Libra-RCNN drop by 16.2%, 40.4%, 51.9%, respectively. Details can be found in *Supplementary Material*.

4.6. Adversarial Defense Methods

We tested five typical methods to defend our attack method in the digital world. These methods included preprocessing defenses (spatial smoothing [40] and Total Variance Minimization [15], adversarial training [14], and their combinations. The most effective way increased the AP from 12.3% to 36.8% only, and our attack method still lowered the AP by 63.2%. See *Supplementary Material* for details.

5. Conclusion and Discussion

Summary. This paper presents a new method of design and manufacturing infrared adversarial clothing. We simulated the process from cloth to clothing in the digital world and then designed the adversarial "QR code" pattern. We manufactured infrared adversarial clothing based on a new material aerogel. Compared with the small bulbs board [42], our adversarial clothing hid from infrared detectors from multiple angles.

Limitations. As mentioned in Section 4.4.2, the adversarial clothing had a significant decrease in the ASR when it was far away from the camera. As mentioned in Section 4.5, it is seen that the adversarial patterns were difficult to attack the transformer-based model, since our adversarial patterns were generated based on the CNN models.

Potential Negative Impact. Adversarial example techniques should be used carefully. If abused, adversarial attacks may threaten the security of AI systems. However, adversarial attack also promotes the research of defense methods.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. U19B2034, 62061136001, 61836014) and the Tsinghua-Toyota Joint Research Fund.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018. 2, 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 5

[3] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE Symposium on Security and Privacy, SP*, 2021. 2

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 2

[5] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS*, pages 15–26, 2017. 2

[6] Tianlong Chen, Yi Wang, Jingyang Zhou, Sijia Liu, Shiyu Chang, Chandrajit Bajaj, and Zhangyang Wang. Can 3d adversarial logos cloak humans? In *CoRR*, 2020. 2

[7] JAS De Lima and J Santos. Generalized stefan-boltzmann law. *International Journal of Theoretical Physics*, 34(1), 1995. 5

[8] Debayan Deb, Jianbang Zhang, and Anil K. Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics, IJCB*, 2020. 2

[9] Ranjie Duan, Xiaofeng Mao, A. Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *CVPR*, 2021. 2

[10] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018. 2

[11] Qi Fan, Ligang Zhang, Honglong Xing, Huan Wang, and Xiaoli Ji. Microwave absorption and infrared stealth performance of reduced graphene oxide-wrapped al flake. *Journal of Materials Science: Materials in Electronics*, 31(4), 2020. 2

[12] FLIR. Free flir thermal dataset for algorithm training. [EB/OL]. https://www.flir.com/oem/adas/adas-dataset-form/ Accessed Nov. 12, 2021. 5

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2, 8

[15] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018. 8

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 8

[17] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *ICCV*, 2021. 2

[18] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, 2020. 2

[19] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 4

[20] Mate Kristo, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using YOLO. *IEEE Access*, 8:125459–125476, 2020. 5

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 2

[22] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *AAAI*, 2019. 2

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 5

[24] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017. 5, 8

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[27] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, 2017. 2

[28] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 5

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6), 2016. 5

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 2

[31] Lei Shang, Yang Lyu, and Wenbo Han. Microstructure and thermal insulation property of silica composite aerogel. *Materials*, 12(6), 2019. 2

[32] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016. 2

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1, 2

[34] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2019. 2

[35] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *CVPR*, 2020. 2

[36] Weilin Wang, Zongwei Tong, Ran Li, Dong Su, and Huiming Ji. Polysiloxane bonded silica aerogel with enhanced thermal insulation and strength. *Materials*, 14(8), 2021. 2

[37] Fred L Bookstein Principal Warps. Thin-plate splines and the decompositions of deformations. *IEEE TPAMI*, 11(6), 1989. 3

[38] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, 2018. 2

[39] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *ECCV*, 2020. 2

[40] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS*, 2018. 8

[41] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, 2019. 2

[42] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *AAAI*, 2021. 1, 2, 5, 8