HOTCOLD Block: Fooling Thermal Infrared Detectors with a Novel Wearable Design

Hui Wei[*1,2], **Zhixiang Wang**[*3,4,5], **Xuemei Jia**[1,2],
Yinqiang Zheng[3], **Hao Tang**[6], **Shin'ichi Satoh**[5,3], **Zheng Wang**[†1,2]

[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering
[3]The University of Tokyo
[4]RIISE
[5]National Institute of Informatics
[6]CVL, ETH Zurich

Abstract

Adversarial attacks on thermal infrared imaging expose the risk of related applications. Estimating the security of these systems is essential for safely deploying them in the real world. In many cases, realizing the attacks in the physical space requires elaborate special perturbations. These solutions are often *impractical* and *attention-grabbing*. To address the need for a physically practical and stealthy adversarial attack, we introduce HOTCOLD Block, a novel physical attack for infrared detectors that hide persons utilizing the wearable Warming Paste and Cooling Paste. By attaching these readily available temperature-controlled materials to the body, HOTCOLD Block evades human eyes efficiently. Moreover, unlike existing methods that build adversarial patches with complex texture and structure features, HOTCOLD Block utilizes an SSP-oriented adversarial optimization algorithm that enables attacks with pure color blocks and explores the influence of size, shape, and position on attack performance. Extensive experimental results in both digital and physical environments demonstrate the performance of our proposed HOTCOLD Block. *Code is available https://github.com/weihui1308/HOTCOLDBlock*.

Introduction

Deep Neural Networks (DNNs) have achieved great success in various fields. They work not only well in visible light but also in thermal infrared imaging, *e.g.*, thermal infrared detection systems that are widely used in autonomous driving, night surveillance, temperature measurement, *etc*. However, adversarial attacks in both *digital* and *physical* worlds expose the vulnerability of DNNs, raising concerns about the security of related applications. The security of DNNs has attracted significant attention in *visible* light (Xu et al. 2021; Duan et al. 2021; Cai et al. 2022; Wang et al. 2022; Hu et al. 2022; Zhong et al. 2022) but has not been fully explored in thermal *infrared* imaging.

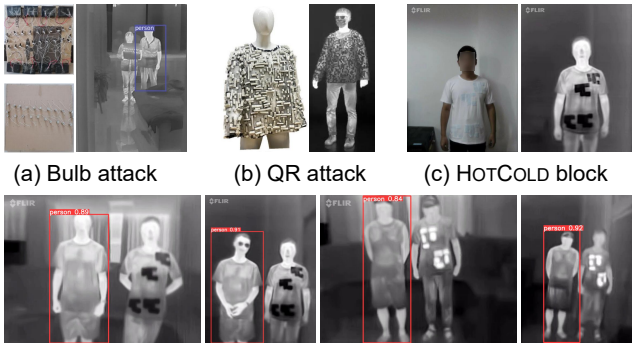

Figure 1: Different infrared attack methods. Our HOTCOLD Block is effective and stealthy. It achieves competitive attack performance on YOLOv5 (Jocher 2020) while evades human eyes better.

This paper focuses on the *physical* adversarial attack on *infrared* detectors (for brevity, we refer to "thermal infrared" as "infrared" throughout the paper), which *hides* persons from smart infrared cameras. Different from adversarial attack in the *digital* world that directly injects adversarial perturbations to captured images, adversarial attack in the *physical* world requires physically realizable objects—defined as adversarial medium—to compose adversarial perturbations. Recently, Zhu et al. (2021, 2022) successively propose to perform attacks on infrared detectors with patches that consists of small light bulbs and invisible clothing made of aerogel. While they achieve reasonable attack effectiveness, their adversarial mediums are attention-grabbing and look unnatural to the human, making the attack suspicious.

To address the aforementioned problems, we introduce new adversarial mediums—Warming Paste and Cooling Paste—based on our findings. First, such temperature-controlled materials can affect infrared imaging. They appear as pure color blocks under the infrared camera and interfere with detectors consequently. Second, they are wearable, making the attacks quite stealthy. Figure 1 gives the

[*]These authors contributed equally.
[†]Corresponding Author

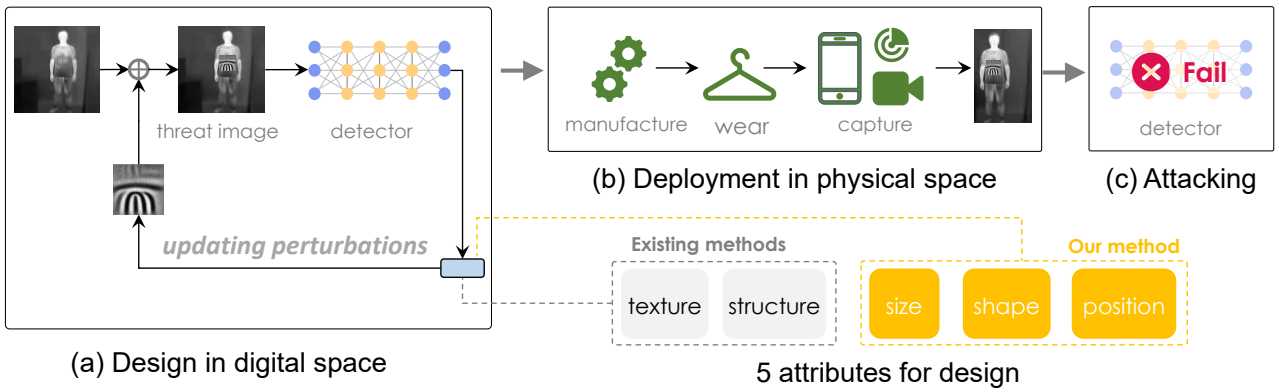

Figure 2: Distinguishing the proposed method with existing methods. Unlike prior methods, the proposed HOTCOLD Block optimizes the size, shape, and position of patches that are pervasively overlooked.

comparison samples. Third, they are physically practical for performing the attacks, only need to paste them on the body. And finally, the Warming Paste and Cooling Paste are readily available without complicated hand-crafting. The cost to execute an attack with them is less than $1. Consequently, they are ideal for attacking infrared detectors, and we pose a new research question: *how to design effective patterns with these new adversarial mediums*?

This research problem is challenging since images of these materials are simple under infrared cameras and can hardly construct complex patterns. Existing physical adversarial attacks (Thys, Van Ranst, and Goedemé 2019; Liu et al. 2020; Tan et al. 2021; Hu et al. 2021) mainly use patch-based methods, which replace a localized region of the targeted image with an elaborate patch. In this regard, researchers delve into the structure and texture features of the patch. They aim to generate particular patterns that can fool the DNNs. Intuitively, the attack effectiveness would suffer from the simple structure and texture of the patch.

Considering this limitation and the imaging characteristics of the Warming Paste and Cooling Paste, we propose HOTCOLD Block, which exploits pure color blocks to achieve physically practical and stealthy adversarial attacks under infrared cameras (Figure 2). To improve the performances in attack, we analyze the 5 attributes of patches: shape, size, position, structure, and texture, and design SSP-oriented adversarial optimization, which optimize the patch's size, shape, and position on the target human body simultaneously. To the best of our knowledge, we are the first to reveal how the lower-level attributes—size, shape, and position—compared to the high-level counterparts—texture and structure—affect the attack performance. Extensive experimental results on the digital and physical world demonstrate that HOTCOLD Block can effectively attack infrared detectors while ensuring quite stealthy.

Our main contributions are summarized below:

- We propose a new stealthy adversarial attack via the wearable temperature-sensitive materials Warming Paste and Cooling Paste, called HOTCOLD Block, which is physically practical.
- We develop an SSP-oriented adversarial optimization that considers three lower-level features of patches simultaneously, *i.e.*, size, shape, and position, instead of setting them manually like most prior works.
- We evaluate our method on mainstream detectors. Extensive experiments in digital and physical space show that our HOTCOLD Block achieves competitive performance on effectiveness, stealthiness, and robustness.

Related Work

Patch-based adversarial attacks. The patch-based adversarial attack is defined as an attack that is able to fool DNNs with elaborate patches and has been frequently applied to physical attacks (Liu et al. 2019a, 2020; Zolfi et al. 2021). Generally, this type of method replaces a localized region of the threat image with a patch, regardless of perturbation constraint (Brown et al. 2017; Liu et al. 2019b; Zhu et al. 2021). In recent years, researchers aimed to trade off the effectiveness and stealthiness of the patch. For instance, Thys, Van Ranst, and Goedemé (2019) managed to generate smoother textures with the total variation loss. After that, Wu et al. (2020) and Xu et al. (2020) printed patches on the clothing to evade human eyes, *e.g.*, adversarial T-shirt and invisibility cloak. Some recent papers generated cartoon-like patches that look more natural (Tan et al. 2021; Hu et al. 2021). To sum up, previous methods mainly focus on designing special structures and textures for adversarial patches while overlooking the more general attributes: size, shape, and position. In this paper, we explore the influence of those three attributes on attack effectiveness.

Attacks to thermal infrared imaging. Unlike the extensive research work on adversarial attacks in visible light images, to the best of our knowledge, only two publications focus on the safety of thermal infrared imaging. Zhu et al. (2021) proposed a patch-based adversarial attack, which uses small glowing light bulbs to manufacture special infrared patterns. The following year, Zhu et al. (2022) designed infrared invisible clothing based on a new material aerogel that successfully evades person detectors. Obviously, whether small glowing light bulbs or clothes made of aerogel material, they have a common shortcoming: they

are attention-grabbing when performing attacks. This is contrary to the mission of adversarial attacks. Unlike those works, we propose a physically practical and stealthy adversarial attack called HOTCOLD Block. The adversarial mediums we use are the Warming Paste and Cooling Paste, wearable and readily available temperature-controlled materials.

Method

This section presents our method, *i.e.*, HOTCOLD Block. We first introduce the modeling of the adversarial medium and then describe the SSP-oriented adversarial optimization.

Problem Definition

Given an input image I and the distribution of all original images $\mathcal{D}$, each $I \in \mathcal{D}$ contains one or multiple person instances. The pre-trained person detector $f : I \rightarrow \mathcal{Y}$ can predict labels $\hat{\mathcal{Y}}$ matching the true labels $\mathcal{Y}$ that includes position of the bounding boxes $\mathcal{V}_{pos}$, the object probability $\mathcal{V}_{obj}$ and the class score $\mathcal{V}_{cls}$:

$$\hat{\mathcal{Y}} := [\mathcal{V}_{pos}, \mathcal{V}_{obj}, \mathcal{V}_{cls}] = f(I). \tag{1}$$

Our goal is to fool the person detector so that it cannot identify the person, *i.e.*, $\mathcal{V}_{obj}$=0. In this paper, we use a patch-based attack method, which replaces localized regions of the original image with patches. We denote the threat image as I_{adv}. The goal can be described as follows:

$$\arg\min \mathcal{V}_{obj} = \arg\min_{i} f(I_{adv}), \tag{2}$$

where i is the index of the i-th image in $\mathcal{D}$.

HOTCOLD Block Modeling

HOTCOLD Block aims to fool the infrared detector using the Warming Paste and Cooling Paste, essentially a patch-based adversarial attack. A patch generally has five attributes: size, shape, position, structure, and texture. In this paper, we delve into the size, shape, and position ("SSP" for short) of patches, with the two primary considerations: *(i)* in the previous work, the lower-level features SSP of patches are pervasively set manually. Their influence on attacks has not been fully explored compared to the high level of features—texture, and structure. *(ii)* since the Warming Paste and Cooling Paste are imaged as pure color blocks under the infrared camera with simple structure and texture, studying the SSP of color blocks would be more meaningful.

In practice, allowing the optimization algorithm to fit an arbitrary shape is unreasonable. One is because some shapes are not physically achievable with our adversarial medium, and the other is due to the high complexity of area calculation for some shapes. These problems are especially prominent for irregular concave polygons. Thus, we use a nine-square-grid to model our adversarial medium in the digital space, as shown in Figure 3. Consequently, not only are the aforementioned problems tackled, but we can simulate the SSP of patches reasonably and conveniently.

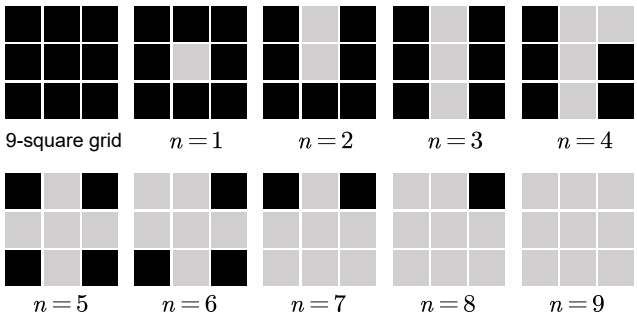

Figure 3: Example of nine-square-grid states for shape modeling. We use the optimization algorithm to find the optimal shape that minimizes the object probability $\mathcal{V}_{obj}$. As shown, the available options are diverse and flexible.

Size. Here, we define the size as the area of patches. Since HOTCOLD Block uses more than one patch to attack the targeted infrared detector, the size depends on the number m of patches, the number n of the occupied grids, and the side length l of the nine-square-grid. Note that the small size facilitates stealthiness. To minimize patch size, we design a mechanism to trade off the size and attack effectiveness. Concretely, we take the growth part of the size as a penalty term, represented as

$$\mathcal{L}_{obj} = \mathcal{V}_{obj} + \underbrace{\lambda \Delta_{\uparrow}(\frac{n(m \times l^2)}{9})}_{\text{penalty term}}, \tag{3}$$

where λ is a hyperparameter to prevent $\mathcal{V}_{obj}$ from overwhelming the penalty term. The sign $\Delta_{\uparrow}$ represents positive growth, and we set the penalty term to 0 when it decreases.

Shape. For the shape of patches, we determine it with a 3×3 matrix $\mathcal{M}$. The 0-1 value of the matrix controls the state of each grid in the nine-square-grid, *e.g.*, $\mathcal{M}$=[[0, 1, 0], [1, 1, 1], [0, 1, 0]] represents the state shown in Figure 3 (n=5). As shown, we can exploit flexible and complex combinations to obtain a plethora of patch shapes. The optimization algorithm in our method is dedicated to finding the optimal shape to achieve a higher attack success rate.

Position. To improve the attack effectiveness, HOTCOLD Block employs a multi-patch joint attack strategy. Our adversarial mediums, the Warming Paste and Cooling Paste, are suitable for this, and it is realizable and natural to achieve in the physical space. Here, we find an appropriate set of coordinates $\mathcal{P}$=$\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ to determine the position of the top-left vertex of each patch, where m is the number of patches. The corresponding vertices coordinate $\mathcal{P}$ is served as the parameter for optimization.

SSP-oriented Adversarial Optimization

In Figure 4, we display the adversarial mediums and their imaging under the infrared camera. Based on the aforementioned modeling, we develop an SSP-oriented adversarial optimization algorithm for performing successful attacks, with the core objectives of **1)** minimizing patch Size, **2)** finding the optimal Shape, **3)** learning suitable Positions.

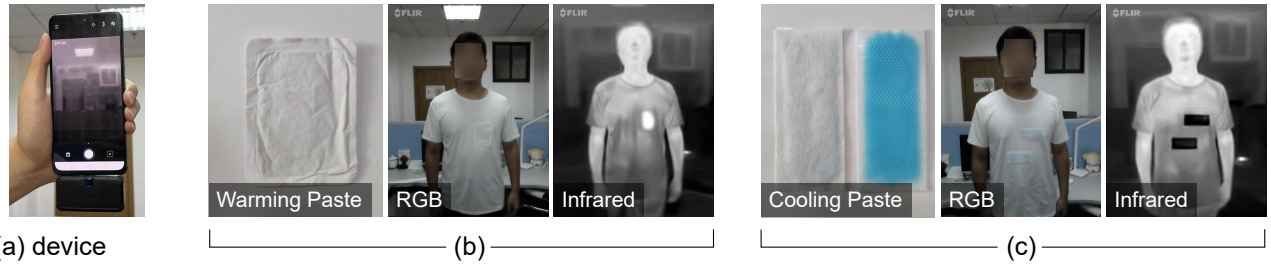

Figure 4: The hardware used for physical attack. (a) our image acquisition device. (b) the Warming Paste with their images in RGB-Infrared space. (c) the Cooling Paste with their images in RGB-Infrared space.

Algorithm 1: SSP-oriented Adversarial Optimization

Input: Dataset $\mathcal{D}$, Detector f.
Parameter: A vector of parameter set $\mathcal{O} = \{\mathcal{M}, \mathcal{P}\}$
Output: $\mathcal{M}$, $\mathcal{P}$

Let $t = 0$.
Initialization: Randomly set $\mathcal{M}$, $\mathcal{P}$
InitializeSwarm(particles)
for $i \leftarrow 0$ to $epoch$ **do**
  while $I = iterator(\mathcal{D})$ is not Null **do**
    $I_{adv} \leftarrow apply(I, \mathcal{M}, \mathcal{P})$
    $\mathcal{L}_{obj} = f(I_{adv})$
    $\mathcal{M}^{'}, \mathcal{P}^{'} \leftarrow$ particles.$move(\vec{x}_i, \vec{v}_i)$
    $I^{'}_{adv} \leftarrow apply(I, \mathcal{M}^{'}, \mathcal{P}^{'})$
    $\mathcal{L}^{'}_{obj} = f(I^{'}_{adv})$
    if $\mathcal{L}^{'}_{obj} < \mathcal{L}_{obj}$ **then**
      $\mathcal{M} \leftarrow \mathcal{M}^{'}, \mathcal{P} \leftarrow \mathcal{P}^{'}$
      particles.$update$(PersonalBest: $\vec{pb}_i$)
      swarm.$update$(GlobalBest: $\vec{gb}_i$)
    else
      pass
    end if
  end while
end for
return GlobalBest: $\{\mathcal{M}, \mathcal{P}\}$

Before describing the optimization algorithm, we analyze the optimization objective and optimization parameters. The focus of the optimization objective is to minimize $\mathcal{L}_{obj}$, aiming to result in person invisibility by detector f. The optimized features are the patch's size, shape, and position. Formally, we list the optimization parameter set $\mathcal{O}$=$\{\mathcal{M}, \mathcal{P}\}$.

Since calculating a backward gradient on all operations of HOTCOLD Block is challenging and the parameter values in $\mathcal{O}$ are discrete, solving this problem with the popular gradient descent optimization algorithm (Ruder 2016) is not appropriate. Inspired by Zhong et al. (2022), we exploit the particle swarm optimization (PSO) strategy (Poli, Kennedy, and Blackwell 2007), which is a bio-inspired algorithm and does not use the gradient of the problem being optimized.

Based on the PSO, we design the SSP-oriented adversarial optimization. Specifically, a number of simple entities, the particles, are placed in the search space of our problem. Each individual in the particle swarm is composed of $\mathcal{O}$=$\{\mathcal{M}, \mathcal{P}\}$. On each iteration, all the particles adjust their velocities $\vec{v}_i$ and positions $\vec{x}_i$. If one position is better than any that has been found so far, then the value is stored as the globe best position $\vec{gb}_i$ of the swarm. Meanwhile, the individual particle has its own personal best position $\vec{pb}_i$. The pseudocode of SSP-oriented Adversarial Optimization is shown in Algorithm 1. We start with a number of random points. All the particles move in the direction of decreasing $\mathcal{L}_{obj}$. Each movement of particles is influenced by the $\vec{pb}_i$ but is also guided toward the $\vec{gb}_i$, which is found by the entire swarm of particles. In our optimization, the number of parameters is few, reduced by 4000 times, compared to updating the patch's structure and texture (a 300×300 patch that has 9×10^4 update pixels).

Experiments

In this section, we carefully evaluate the performance of our HOTCOLD Block on the three criteria: *effectiveness*, *stealthiness*, and *robustness*.

Experimental Settings

Datasets. **1)** Digital adversarial attack. Following Zhu et al. (2021), we evaluate the performance of our method on the Teledyne FLIR ADAS Thermal dataset (SYSTEMS 2022b)[1]. Infrared images were acquired with a Teledyne FLIR Tau2 (13 mm f/1.0 with a 45-degree HFOV and 37-degree VFOV). The thermal camera operated in T-linear mode. We filter the original dataset for better fitting to the patch-based adversarial attack, with two conditions of *(i)* the images contain "person" category, *(ii)* the bodies of persons in the images have a height of more than 120 pixels. Finally, 1,255 images are available, of which 878 are the training set with 1,366 eligible "person" labels and 377 are the testing set with 598 eligible "person" labels. **2)** Physical adversarial attack. In the physical space, we capture infrared images with a FLIR ONE Pro camera (SYSTEMS 2022a), which has a thermal resolution of 160×120. During capture, the camera is connected to a Xiaomi phone for real-time image display (see Figure 4(a)). We build lab setups that allow for shooting distances from 0 to 4 meters, record 8 videos in

[1]Note that since the Teledyne FLIR company released an updated version in January 2022, we use v2.0 instead of v1.0. The updated dataset not only expands labels to 15 categories *vs.* 5 original categories, but also the scale of the annotated image with an +83% increase compared to the v1.0 release.

Num of Patches m	Method	Side length l (%)													
		6		8		10		12		14		16		Average	
		AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑
1	R	94.1	1.3	94.4	0.7	93.4	0.0	93.3	2.2	93.0	2.9	91.1	7.0	93.2	2.4
	MR	94.7	0.0	94.2	4.0	94.3	1.8	93.7	3.5	94.3	2.2	93.7	2.6	94.2	2.4
	HCB	93.4	1.8	93.2	2.9	90.7	3.7	88.0	8.4	85.6	9.0	80.9	14.7	**88.6**	**6.8**
2	R	93.4	3.7	92.9	3.7	92.6	1.7	91.0	5.0	87.4	12.7	81.8	18.5	89.9	7.6
	MR	93.9	2.6	93.2	2.8	92.5	2.2	92.2	5.3	90.5	6.4	75.0	24.9	89.6	7.4
	HCB	92.0	5.9	87.4	7.5	81.0	21.5	70.3	26.8	65.3	29.2	54.8	40.9	**75.1**	**22.0**
3	R	93.3	0.0	92.2	1.3	91.2	5.5	86.5	14.9	82.5	15.4	76.6	21.3	87.1	9.7
	MR	94.0	3.1	93.3	2.2	86.4	13.0	79.0	19.4	70.0	33.0	58.8	35.8	80.3	17.8
	HCB	89.4	6.6	86.0	10.8	70.4	28.6	58.7	33.8	52.0	37.8	38.3	49.2	**65.8**	**27.8**
4	R	91.5	3.1	89.7	4.6	87.6	11.6	83.7	17.8	73.8	25.9	66.1	35.9	82.1	16.5
	MR	93.5	3.5	92.3	4.8	81.8	19.3	74.7	23.9	65.8	37.2	49.0	38.7	76.2	21.2
	HCB	81.6	13.2	69.5	25.0	66.6	31.9	43.0	40.4	26.6	59.3	29.3	59.8	**52.8**	**38.3**
5	R	92.1	1.1	87.6	10.1	81.7	19.3	72.5	26.1	63.7	36.1	47.2	46.8	74.1	23.3
	MR	91.8	2.9	86.8	9.9	64.9	31.4	48.3	47.9	28.2	69.4	14.1	77.8	55.7	40.0
	HCB	83.5	13.6	58.4	38.3	49.8	47.2	34.6	58.7	22.9	68.8	11.6	77.8	**43.5**	**50.7**
6	R	90.7	7.5	87.6	10.1	78.4	19.1	64.0	33.0	62.0	39.3	31.0	59.9	69.0	28.2
	MR	90.9	7.9	84.9	13.4	76.8	19.8	44.8	50.1	25.9	67.3	8.9	82.6	55.4	40.2
	HCB	75.3	22.9	59.7	39.3	40.1	57.4	35.7	60.9	11.0	82.2	10.2	83.5	**38.7**	**57.7**
7	R	88.2	8.8	81.9	15.8	68.4	29.7	59.4	37.9	48.4	45.1	25.2	64.6	61.9	33.7
	MR	88.6	13.6	83.0	14.3	69.4	25.7	28.5	64.2	14.5	73.0	5.7	88.4	48.3	46.5
	HCB	71.9	26.2	57.4	34.5	36.1	56.9	24.0	66.8	24.8	71.0	5.9	89.9	**36.7**	**57.6**

Table 1: Quantitative results on the FLIR ADAS test set at varying setups. We report AP (%), ASR (%) for our adversarial attack method HOTCOLD Block (HCB) *vs.* the random block attack (R) and the manual-random block attack (MR), under varying numbers m of patches and side lengths l (% of the person's height).

different scenes, and then extract one frame per second. We capture a total of 112 images and use LabelImg[2] to annotate them with 224 labels.

Evaluation metrics. We aim to hide the person from detectors. To this end, we adopt the Average Precision (AP) metric to evaluate the performance of detectors on the threat dataset. Note that lower AP indicates stronger attack effect. In addition, the Attack Success Rate (ASR) is used to evaluate the effectiveness of our attack methods, which is defined as the percentage of positive and total samples as follows:

$$\text{ASR}(X) = 1 - \frac{1}{N}\sum_{i=0}^{N} \text{sign}(label_i), \tag{4}$$

$$\text{sign}(label_i) = \begin{cases} 1, & label_i \in L_{pre} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where N is the number of all true positive labels detected in the dataset X when there is no attack, L_{pre} is the set of all labels detected under attacking. The higher the ASR, the more effective the adversarial attack method is.

Competing methods. We compare our method with the only 2 methods in the field of infrared attack:

- **Bulb Attack** (Zhu et al. 2021): a physical attack method that fools infrared pedestrian detectors using small bulbs.
- **QR Attack** (Zhu et al. 2022): a multi-angle physical attack method that designs the adversarial "QR code" pattern for attacking infrared detectors.

[2]LabelImg: https://github.com/heartexlabs/labelImg

Implementation details. We use YOLOv5 (Jocher 2020) as our target model since it is a fast, effective, and widely-used detector. For infrared detection, we use the pre-trained weights on the MSCOCO Dataset (Lin et al. 2014) as initialization and fine-tune the model on the FLIR ADAS Dataset. The AP score of the fine-tuned target model achieves 94.8% on the test set. We set the population size of the particle swarm to 100 and the block's pixel value to 0.2. We conduct all experiments on a device with an NVIDIA GeForce RTX 3090 GPU, and all our codes are implemented in PyTorch. Our used Warming Paste can keep on heating for 6 hours with an average temperature of 53 ℃, and Cooling Paste can cool down to 24 ℃ for 4 hours.

Evaluation of Effectiveness

Digital adversarial attack. In the digital space, we attack every image in the test set of the FLIR ADAS dataset with HOTCOLD Block. We run our attack with controlled numbers of patches and side lengths. Table 1 reports the effectiveness evaluation results for our method (HCB) *vs.* the random block attack (R) and the manual-random block attack (MR). MR refers to random positions, but manual intervention to avoid overlap between patches for fair comparison. Through the analysis of all experimental setups and results in Table 1, we can reach the following three conclusions: **1)** HOTCOLD Block comprehensively outperforms random and manual-random, demonstrating that our method is feasible and effective; **2)** The attack effectiveness is strengthening as the number m of patches and the side length l increase, in keeping with our expectations; **3)** HOTCOLD Block lowers the AP to 43.0% and achieves 40.4% ASR under m=4 and

(a) Clean image (b) Random patch (c) Bulb Attack (d) QR Attack (e) HOTCOLD Block

Figure 5: Example results of digital attacks. The bounding boxes indicate the infrared detector successfully detects the person.

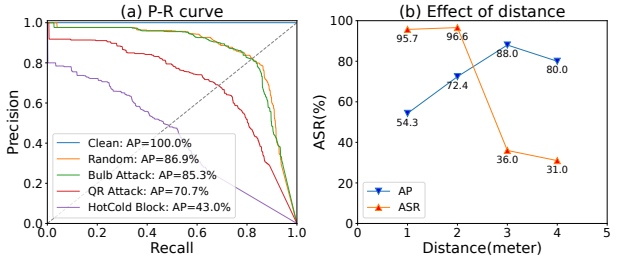

Figure 6: Quantitative results. (a) Precision-Recall curve in the digital space. (b) The AP(%) and ASR(%) of HOTCOLD Block at different distances in the physical space.

l=12, showing that our method is physically practical due to the legitimate configuration. Note that our following experiments are in this configuration unless otherwise specified.

In Figure 5, we show qualitative examples comparing our results with those of the baseline methods. Moreover, we draw the P-R curve for quantitative evaluation in Figure 6(a). It is clear that HOTCOLD Block achieves competitive performance. For example, it causes the AP of YOLOv5 to drop by 51.8%, significantly outperforming 9.5% and 24.1% achieved by the only two baseline methods. Note that, although we applied four patches for attacking (as shown in Figure 5(e)), the total area is similar to the area of 1 patch applied in the comparison methods.

Physical adversarial attack. Figure 6(b) and Figures 8 depict the quantitative and qualitative results, respectively. Our attack is effective and achieves an over 90% ASR at 1m and 2m shooting distances. Although only attacking one person in the frame, the AP drops to 73.6% on average. Observe that with the increasing shooting distance, the ASR becomes relatively low. By comparison, the shape of the patch suffers from obvious deformation due to the insufficient infrared camera resolution. This deformation may degrade the attack effectiveness. Nevertheless, the ASR is still around 34%. The results show that HOTCOLD Block is physically practical and evades infrared detectors in the real world. See *Supplementary Material* for the video demo.

Evaluation of Stealthiness

As aforementioned, considering the stealthiness of attacks, we choose the Warming Paste and Cooling Paste as our adversarial medium. In Figure 7(a), it is clear that the adversarial mediums are in harmony with society and do not draw attention to themselves. Moreover, even if the status quo is still unsatisfactory, our method allows hiding the adversarial medium by wearing another garment on the outer surface. As shown in Figure 7(b), the human eyes cannot recognize adversarial mediums on the human body in RGB space. Note that the change from the "exposed" to the "covered" hardly affects the infrared camera's imaging. Compared with the baseline methods, HOTCOLD Block successfully achieves fairly imperceptible attacks.

(a) Exposed (b) Covered

Figure 7: Examples of RGB-Infrared image pairs with exposed and covered adversarial mediums.

Detector	*w/o* Attack		*w/* Attack	
	AP↓	ASR↑	AP↓	ASR↑
YOLOv3	95.4	–	52.5	40.1
YOLOv5	94.8	–	43.0	40.4
DETR	94.2	–	65.1	4.5
RetinaNet	93.6	–	57.5	22.6
Faster RCNN	94.5	–	31.9	49.4
Mask RCNN	95.7	–	48.8	36.3
Average	94.7	–	49.8	32.2

Table 2: Evaluation across various detectors.

Evaluation of Robustness

We evaluate the attack robustness of our approach across various detectors under the black box setting, including YOLOv3 (Redmon and Farhadi 2018), DETR (Carion et al. 2020), RetinaNet (Lin et al. 2017), Faster RCNN (Ren et al. 2016), and Mask RCNN (He et al. 2017). The detectors are pre-trained on the MSCOCO Dataset (Lin et al. 2014) and fine-tuned on the FLIR ADAS Dataset. Table 2 reports ASR and the changes in AP. We can see that the detectors have a significant degradation in performance when facing HOTCOLD Block, which makes AP drop by 44.9% and achieves 32.2% in ASR on average. By comparing the results, we also notice that DETR is minimally affected. One possible reason is that the transformer-based network used in DETR is beneficial in defending against adversarial attacks. Broadly, our attack is shown to be robust under the majority of models.

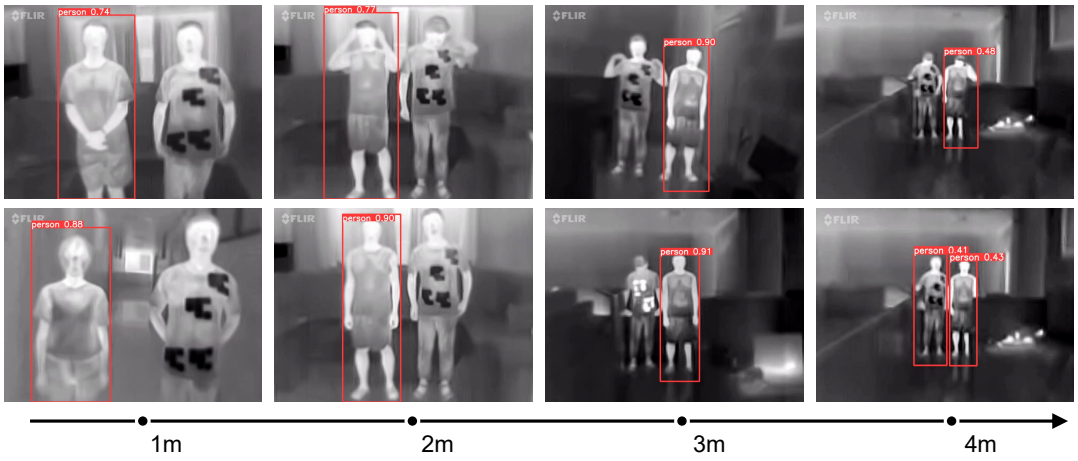

Figure 8: Example results of physical attacks for HOTCOLD Block at different distances.

Parameter and Ablation Studies

Effect of λ. Larger λ have more effect on stealthiness by reducing the size but are less effective. We evaluate the AP and ASR of our attack with different λ. As shown in Figure 9(a), the attack effectiveness of HOTCOLD Block remains stable when λ equals 0 to 3, and drops dramatically when λ=4. Moreover, We analyze the state of the nine-square-grid and observe that as λ increases, fewer grids are activated, *i.e.*, the actual area of the patch is smaller. To sum up, by modifying λ, our proposed method can trade off the visual stealthiness and effectiveness, and we set λ=3 to obtain the optimal balance.

Effect of the block's pixel value. We then show how the block's pixel value impacts the adversarial effectiveness of HOTCOLD Block. Figure 9(b) shows the results. Observe that as the pixel value increases, the attack capability exhibits a concave curve, which means that taking values close to 0 or 1 favors the attack. This change is because when the patch and the body are fused, *i.e.*, their pixel values are close, and the patch no longer has attack capabilities. Note that the values of the Warming Paste and Cooling Paste under the infrared camera are nearly 0.9 and 0.2, respectively. Therefore, the adversarial mediums we chose are appropriate for efficient physical attacks.

Defense Discussion

Attack and defense develop parallel, like an arms race, to improve the model's capabilities. Here, we discuss the method for defending against HOTCOLD Block. Based on our understanding of HOTCOLD Block, we apply the adversarial training (Bai et al. 2021) that aims to enhance the robustness of models intrinsically. Concretely, we augment training data with adversarial examples generated by HOTCOLD Block in each training loop. Then we perform attacks on the retrained model to verify its effectiveness in the digital space. By comparing the results, where the AP achieves 94.9% with no attack and 91.6% with attacks, the retrained model becomes more robust with no performance loss. Therefore, we can use our image augmentation method

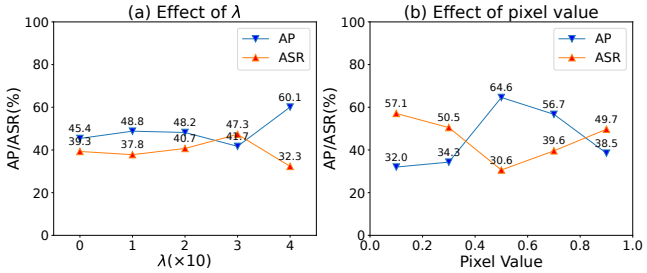

Figure 9: Parameter and ablation studies of hyperparameter λ (a) and the block's pixel value (b).

to improve the detectors' performance further. In this regard, our work has important practical significance for applying DNNs-based models in the real world.

Limitation

Although HOTCOLD Block shows excellent performance in attacking infrared detectors, it is hard to implement multi-angle attacks due to the *segment-missing* problem (Hu et al. 2022). Concretely, since the blocks produced by the Warming Paste and Cooling Paste would be obscured when the viewing angle changes, causing them to disappear from the infrared camera, HOTCOLD Block would drop in the attack success rate. We believe simulating the 3D virtual human body and sticking adversarial blocks on the surface can solve this problem.

Conclusion

In this paper, we propose a novel physical adversarial attack called HOTCOLD Block that applies the Warming Paste and Cooling Paste to hide persons from being detected by infrared detectors. Our wearable adversarial mediums are physically practical and stealthy due to their intrinsic properties. Moreover, we design an SSP-oriented adversarial optimization, which delves into the feature space of size, shape, and position rather than texture and structure. Extensive experiments in both digital and physical spaces show that

our HOTCOLD Block evade both human eyes and detection models more effectively than existing methods.

Ethics Statement

Our work successfully achieves physical adversarial attacks and illustrates the vulnerability of deep learning-based models. As the medium of adversarial attack is readily available and operational, most existing deep learning applications have potential security threats. Our work is dedicated to providing motivation and insights for better defense against malicious attacks.

Acknowledgements

This research is supported by the National Key R&D Project (2021YFC3320301), National Natural Science Foundation of China (62171325), Hubei Key R&D Project (2022BAA033), the CAAI-Huawei MindSpore Open Fund, JSPS KAKENHI Grant Numbers 22H00529, JP22H03620, and the Value Exchange Engineering, a joint research project between Mercari, Inc., and RIISE. Zhixiang thanks to the MEXT Scholarship.

References

Bai, T.; Luo, J.; Zhao, J.; Wen, B.; and Wang, Q. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. In *International Joint Conference on Artificial Intelligence*. Survey Track.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Cai, Z.; Xie, X.; Li, S.; Yin, M.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Asif, M. S. 2022. Context-aware transfer attacks for object detection. In *AAAI Conference on Artificial Intelligence*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European conference on computer vision*.

Duan, R.; Mao, X.; Qin, A. K.; Chen, Y.; Ye, S.; He, Y.; and Yang, Y. 2021. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*.

Hu, Y.-C.-T.; Kung, B.-H.; Tan, D. S.; Chen, J.-C.; Hua, K.-L.; and Cheng, W.-H. 2021. Naturalistic Physical Adversarial Patch for Object Detectors. In *IEEE/CVF International Conference on Computer Vision*.

Hu, Z.; Huang, S.; Zhu, X.; Sun, F.; Zhang, B.; and Hu, X. 2022. Adversarial Texture for Fooling Person Detectors in the Physical World. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jocher, G. 2020. YOLOv5 Detector. https://github.com/ultralytics/yolov5. Accessed: 2023-03-13.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*.

Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; and Tao, D. 2019a. Perceptual-sensitive gan for generating adversarial patches. In *AAAI conference on artificial intelligence*.

Liu, A.; Wang, J.; Liu, X.; Cao, B.; Zhang, C.; and Yu, H. 2020. Bias-based universal adversarial patch attack for automatic check-out. In *European Conference on Computer Vision*.

Liu, X.; Yang, H.; Liu, Z.; Song, L.; Chen, Y.; and Li, H. 2019b. DPATCH: An Adversarial Patch Attack on Object Detectors. In *AAAI Conference on Artificial Intelligence Workshop on Artificial Intelligence Safety*.

Poli, R.; Kennedy, J.; and Blackwell, T. 2007. Particle swarm optimization. *Swarm intelligence*, 1(1): 33–57.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

SYSTEMS, F. 2022a. FLIR ONE Pro Camera. https://www.flir.com/products/flir-one-pro/. Accessed: 2023-03-13.

SYSTEMS, F. 2022b. Teledyne FLIR Free ADAS Thermal Datasets v2. https://adas-dataset-v2.flirconservator.com/. Accessed: 2023-03-13.

Tan, J.; Ji, N.; Xie, H.; and Xiang, X. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In *ACM International Conference on Multimedia*.

Thys, S.; Van Ranst, W.; and Goedemé, T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Wang, D.; Jiang, T.; Sun, J.; Zhou, W.; Gong, Z.; Zhang, X.; Yao, W.; and Chen, X. 2022. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *AAAI Conference on Artificial Intelligence*.

Wu, Z.; Lim, S.-N.; Davis, L. S.; and Goldstein, T. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*.

Xu, K.; Zhang, G.; Liu, S.; Fan, Q.; Sun, M.; Chen, H.; Chen, P.-Y.; Wang, Y.; and Lin, X. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*.

Xu, Q.; Tao, G.; Cheng, S.; and Zhang, X. 2021. Towards feature space adversarial attack by style perturbation. In *AAAI Conference on Artificial Intelligence*.

Zhong, Y.; Liu, X.; Zhai, D.; Jiang, J.; and Ji, X. 2022. Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhu, X.; Hu, Z.; Huang, S.; Li, J.; and Hu, X. 2022. Infrared Invisible Clothing: Hiding from Infrared Detectors at Multiple Angles in Real World. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, X.; Li, X.; Li, J.; Wang, Z.; and Hu, X. 2021. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *AAAI Conference on Artificial Intelligence*.

Zolfi, A.; Kravchik, M.; Elovici, Y.; and Shabtai, A. 2021. The Translucent Patch: A Physical and Universal Attack on Object Detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.