AdvDisplay: Adversarial Display Assembled by Thermoelectric Cooler for Fooling Thermal Infrared Detectors

Hao Li[1,4], **Fanggao Wan**[1,4], **Yue Su**[2], **Yue Wu**[3, 4], **Mingyang Zhang**[1,4], **Maoguo Gong**[1,4*]

[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]School of Artificial Intelligence, Xidian University, Xi'an 710071, China
[3]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[4]Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xi'an 710071, China
haoli@xidian.edu.cn, 22021211895@stu.xidian.edu.cn, 22009200743@stu.xidian.edu.cn, ywu@xidian.edu.cn, myzhang@xidian.edu.cn, gong@ieee.org

Abstract

When the current physical adversarial patches cannot deceive thermal infrared detectors, the existing techniques implement adversarial attacks from scratch, such as digital patch generation, material production, and physical deployment. Besides, it is difficult to finely regulate infrared radiation. To address these issues, this paper designs an adversarial thermal display (*AdvDisplay*) by assembling thermoelectric coolers (TECs) as an array. Specifically, to reduce the gap between patches in the physical and digital worlds and decrease the power of *AdvDisplay* device, heat transfer loss and electric power loss are designed to guide the patch optimization. In addition, a precise temperature control scheme for *AdvDisplay* is proposed based on proportional-integral-derivative (PID) control. Due to the accurate temperature regulation and the reusability of *AdvDisplay* , our method is able to improve the attack success rate and the efficiency of physical deployments. Extensive experimental results indicate that the proposed method possesses superior adversarial effectiveness compared to other methods and demonstrates strong robustness in physical attacks.

Introduction

Thermal infrared imaging is widely used in military, security surveillance, and autonomous driving due to its unique night vision capability and insensitivity to environmental lighting (Wilson et al. 2023). Deep Neural Networks (DNNs), as a powerful feature learning tool (Li et al. 2022; Gong et al. 2023; Gong, Yuan, and Bao 2021; Liu and Tsang 2015), have significantly advanced the development and application of infrared object detection (Zhang and Demiris 2023; Bustos et al. 2023; Kou et al. 2023; Yang et al. 2024). In recent years, adversarial sample attacks that can mislead DNNs have attracted widespread attention (Szegedy et al. 2013; Yuan et al. 2019). These attacks can cause DNNs to output erroneous results with high confidence by adding meticulously designed perturbations to the input data. Furthermore, some related studies have shown that adversarial samples not only exist in the digital world but can also be realized physically, causing challenges to the security of real-world applications (Hu et al. 2022; Ma et al. 2023; Wu et al. 2021; Xiong et al. 2021; Liu and Tsang 2017; Huang et al. 2024).

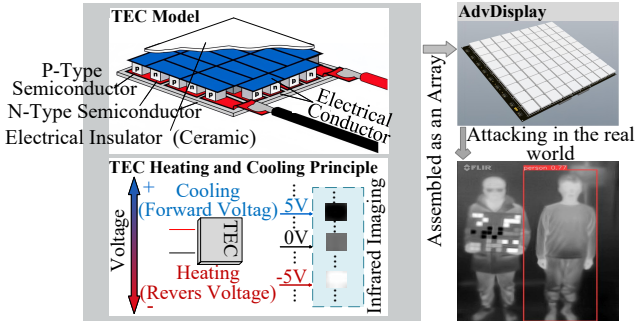

Figure 1: *AdvDisplay* model and visual examples of physical attacks with *AdvDisplay* .

However, the majority of existing research on adversarial sample attacks has primarily concentrated on the visible light domain, with relatively less attention given to infrared adversarial attacks. Additionally, traditional adversarial implementation methods based on RGB appearance are ineffective in the infrared domain due to infrared imaging relying on thermal radiation from the object (Zhu et al. 2021; Wei et al. 2023a). This challenge necessitates researchers to develop novel methodologies for implementing effective physical adversarial attacks in the infrared domain.

Currently, the physical implementation of infrared adversarial attacks can be mainly categorized into active and passive approaches[1]. To our knowledge, the existing active physical adversarial methods are primarily implemented based on small bulbs and resistors (Zhu et al. 2021; Bendelac et al. 2021; Zhu et al. 2023, 2024). However, due to the limitations of the materials, they do not have a high degree of freedom in the regulation of radiation. Passive physical attacks mainly include hot-cold block-based attacks, insulation material-based attacks and others (Wei et al. 2023a; Hu et al. 2024; Zhu et al. 2022; Kim, Lee, and Ro 2022; Gong, Yuan, and Bao 2023; Kim, Yu, and Ro 2023). Hot-

*Corresponding author.

[1]The active approach is based on an external power source to change the infrared characteristics of objects. In contrast, the passive approach implements attacks using a carrier without power.

Related Works	Materials	Description	Precise and Continuous Controllability of Radiation	Bidirectional Adjustability of Radiation	Reusability of Physical Carriers
(Zhu et al. 2022)	Aerogel	A clothing that can achieve multi-angle attack.	Binary radiation control(×)	Insulating heat (×)	N/A (×)
(Wei, Yu, and Huang 2023)	Aerogel	An attack method with learnable shape and position.	Binary radiation control(×)	Insulating heat (×)	N/A (×)
(Wei et al. 2023a)	Hot-cold Block	Hiding people with the heating and cooling pastes.	Binary temperature (×)	Heating and cooling (✓)	N/A (×)
(Hu et al. 2024)	Hot-cold Block	Executing multi-view attack based on hot-cold block.	Binary temperature (×)	Heating and cooling (✓)	N/A (×)
(Zhu et al. 2021)	Small Bulbs	Utilizing small bulbs to implement physical attack.	No temperature feedback (×)	Heating (×)	N/A (×)
(Bendelac et al. 2021)	Resistors	A resistor-based adversarial board is designed.	No temperature feedback (×)	Heating (×)	Array form (✓)
(Kim, Yu, and Ro 2023)	Low-e films	Using Low-e films against multispectral detectors.	Ternary radiation control (×)	Multiple emissivity (✓)	N/A (×)
Ours method	TEC	An adversarial thermal display *AdvDisplay* is designed by assembling TECs as an array.	Electrical control with feedback (✓)	Heating and cooling (✓)	Array form (✓)

Table 1: Related work on infrared adversarial attacks.

cold block-based attacks change the object's infrared characteristics by heating and cooling paste to implement physical attacks. However, since the heating and cooling paste cannot maintain their thermal properties over extended periods and only two radiation intensities can be realized, the hot-cold block-based methods have limitations regarding adversarial effectiveness and stability. The insulation-based adversarial attack methods can currently only change the radiation on specific areas with binary or ternary patterns, which limits the search space for adversarial patches. Overall, a high-attack and low-cost method for infrared physical adversarial attacks should meet three characteristics: *precise and continuous controllability of radiation*, *bidirectional adjustability (satisfying both increase and decrease) of radiation*, and *reusability of physical adversarial carriers*. The precise controllability and bidirectional adjustability of infrared radiation can expand the search space of adversarial patterns, enabling more precise and stable physical implementations, thus improving the threat of the adversarial samples in the physical world. The reusability of physical adversarial carriers can shorten the production cycle of patches, improve efficiency, and reduce costs. However, the existing works have not simultaneously considered these three aspects.

To satisfy the three characteristics mentioned above, this paper designs an adversarial thermal display (*AdvDisplay*) by assembling thermoelectric coolers (TECs) as an array. Based on the characteristics of TEC materials, a PID-based control scheme is proposed to regulate the infrared radiation of *AdvDisplay* accurately, enabling our method to create complex and fine-grained adversarial patterns. In addition, in order to improve the physical implementation and reduce the adversarial cost, heat transfer, and electric power loss are designed to guide the optimization of the patches. It is noted that *AdvDisplay* is a reusable physical adversary carrier, i.e., when the deployed infrared adversary patch does not work, it can generate the target patch without physical remaking, which improves efficiency and reduces cost simultaneously.

Our main contributions are summarized as follows:

1. We are the first to employ the TEC material for infrared adversarial attacks. In this paper, a reusable adversarial device *AdvDisplay* , which can control infrared radiation precisely, continuously, and bidirectionally, is designed and fabricated.
2. According to the characteristics of *AdvDisplay* , heat transfer loss and power loss are designed to guide the optimization, and an accurate temperature control strategy is proposed based on PID technology.

Related Works

To implement adversarial attacks in the infrared domain, Zhu *et al.* (Zhu et al. 2021) creatively proposed the first physical method to attack infrared thermal imaging detectors with small bulbs. However, glowing bulbs are easily found in reality,which makes this attack method lack of stealth. After that, there is a focus on attacks with a stealth patch. Zhu *et al.* (Zhu et al. 2022) used aerogel to create an invisibility clothing with a QR code pattern. On this basis, Wei *et al.* (Wei, Yu, and Huang 2023) fabricated an adversarial infrared patch utilizing aerogel material to achieve better attack effect by searching the position and shape of the patch. These works based on thermal insulation materials are proven to achieve good results across different objects (Wei, Yu, and Huang 2023) and modalities (Wei et al. 2023b). In addition, Wei *et al.* (Wei et al. 2023a) employed wearable heated and cooled pastes to fool infrared detectors and to enhance attack performance by optimizing size, shape, and position. Moreover, Hu *et al.* (Hu et al. 2024) developed a method called AdvIB, which utilizes hot and cold blocks as physical perturbations and employs differential evolutionary optimization to determine the most adversarial physical parameters such as position and angle. In parallel, Kim *et al.* (Kim, Lee, and Ro 2022; Kim, Yu, and Ro 2023) explored attacks with varying thermal intensities by utilizing the spectral properties of materials to design adversarial patches and coatings.

While these efforts have made some good progress, none of them have been able to simultaneously achieve precise and continuous controllability of radiation, bidirectional adjustability of radiation and reusability of physical adversarial carriers. In other words, the above most patches to only set the intensity of infrared radiation within the "high" and "low" binary modes, which limits their attack performance. Additionally, when deployed thermal infrared adversarial patches cannot deceive thermal infrared detectors, they need implement adversarial attacks from scratch, such as digital patch generation, material production, and physical deployment, which results in additional economic and time costs.

Methodology

Modeling and Physical Design of *AdvDisplay*

In the real world, the design of infrared adversarial patch patterns is not as flexible as in the case of visible light adversarial scenarios. On the one hand, this is due to some shapes that are physically impossible to realize with our adversarial

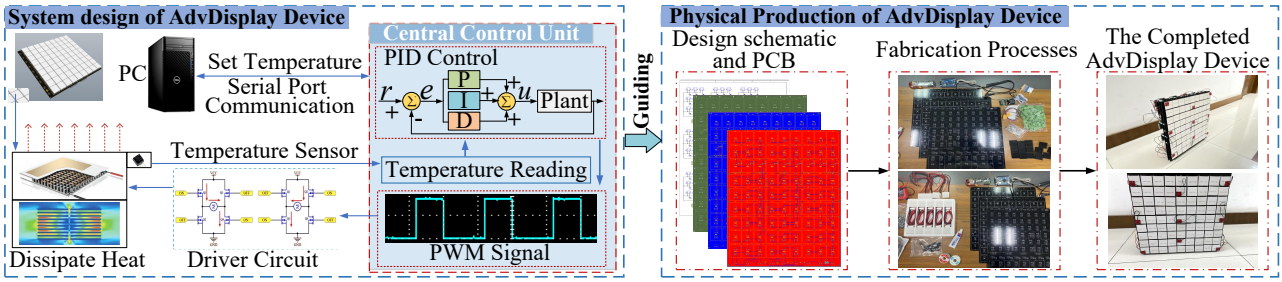

Figure 2: The design and production of AdvDisplay device.

medium; on the other hand, complex shapes such as concave polygons used in adversarial patches often lack reusability. Therefore, the proposed *AdvDisplay* is modeled as an array $\boldsymbol{M}$ of TECs, and then the radiation intensity of each TEC unit on this array is optimized to form a highly adversarial patch. When the physical adversarial patch cannot deceive the detector, a new patch can be created by simply resetting the relevant parameters of the *AdvDisplay* . The physical model of *AdvDisplay* is shown in Figure 1.

In the design of *AdvDisplay* device, a closed-loop feedback control strategy is employed, combined with advanced control algorithms, to achieve high-precision temperature regulation. As illustrated in Figure 2, the system framework primarily comprises four key components: the temperature detection unit, the central control unit, the TEC drive unit, and the user interaction unit.

Firstly, the temperature detection unit employs high-precision temperature sensors to monitor the temperature of the TEC in real-time. The output from these sensors is sent to the central control unit, which contains a microprocessor for executing the temperature control algorithm. The central control unit compares the collected real-time temperature with the reference temperature required for control. Based on the designed temperature control algorithm, it calculates the control quantity and outputs a pulse width modulation (PWM) signal. Subsequently, this PWM is fed into an H-bridge driver circuit to control the heating or cooling of the TEC module, enabling the temperature to rapidly and precisely stabilize at the set value. Additionally, the central control unit can communicate with a personal computer via a serial port to set the working parameters of the physical adversarial board.

In the current market, commonly available sizes of TECs include dimensions such as 1cm $\times$ 1cm, 2cm $\times$ 2cm, and 4cm $\times$ 4cm. A TEC with a size of 4cm $\times$ 4cm are selected for the construction of the physical board, which is analyzed in Section 4.2. Moreover, the microprocessor employed is the STM32F407, which is based on the ARM Cortex-M4 core. It features DSP instructions and a floating point unit, with a maximum clock frequency of up to 168MHz. Finally, to ensure the stability and reliability of the physical adversarial board, the designed circuit and related components are integrated onto a single Printed Circuit Board (PCB).

Attacks in the Digital World

Problem Formulation Assume that I and D denote a clean image and the original dataset, where $I \in D$. A threat image I_{adv} is created by adding an adversarial patch δ to I. Let f and θ represent the detection model and model parameters, respectively. $f(I_{adv}, \theta)$ is defined as the output of the model given the input I_{adv}. Most object detectors have three outputs: position of the bounding boxes f_{pos}, the object probability f_{obj}, and the class score f_{cls}, which can be written as

$$y_{adv} = [f_{pos}, f_{obj}, f_{cls}] = f(I_{adv}, \theta). \tag{1}$$

The goal of the adversarial patch δ is to make pedestrians evade detection by the object detector. In other words, V_{obj} is expected to be as reduced as possible. Assume that there are N pedestrians in D. Therefore, our goal can be described as

$$\min \mathcal{L}_{obj} = \min \frac{1}{N} \sum_{i=1}^{N} f_{obj}^{i}(I_{adv}, \theta). \tag{2}$$

Different from RGB-based adversarial patches, a physical phenomenon known as heat transfer exists between each TEC unit in *AdvDisplay* . This phenomenon blurs the boundary between neighboring TEC units, thus making it challenging to accurately map digital patches to the physical world. In order to decrease the gap between the adversarial patches in the physical world and the digital world, the heat transfer loss $\mathcal{L}_{ht}$ is employed to guide the generation of patches. Since the heat transfer phenomenon between objects with a more considerable temperature difference is more significant, $\mathcal{L}_{ht}$ reduces the total heat transfer by decreasing the temperature difference between neighboring TEC units. $\mathcal{L}_{ht}$ can be expressed as

$$\mathcal{L}_{ht} = \sum_{i,j} \sqrt{(T_{i,j} - T_{i+1,j})^2 + (T_{i,j} - T_{i,j+1})^2}, \tag{3}$$

where $T_{i,j}$ represents the temperature of the TEC with coordinates (i, j) in *AdvDisplay* .

The *AdvDisplay* requires electrical power to maintain its stable infrared adversarial performance. To save electricity cost, the electric power loss $\mathcal{L}_{ep}$ is proposed. It is assumed that the environment temperature is T_{temp}, and T_{obj} is the expected temperature of a TEC unit. In order to keep the temperature stable, the power required to convert electrical energy into thermal energy should be balanced with the heat loss of the TEC unit. Specifically, heat loss is usually caused in three ways: conduction, convection, and radiation (Sidebotham 2015). Generally, heat loss due to radiation is only considered in high-temperature applications, so conduction and convection are the main ways of heat loss in TECs.

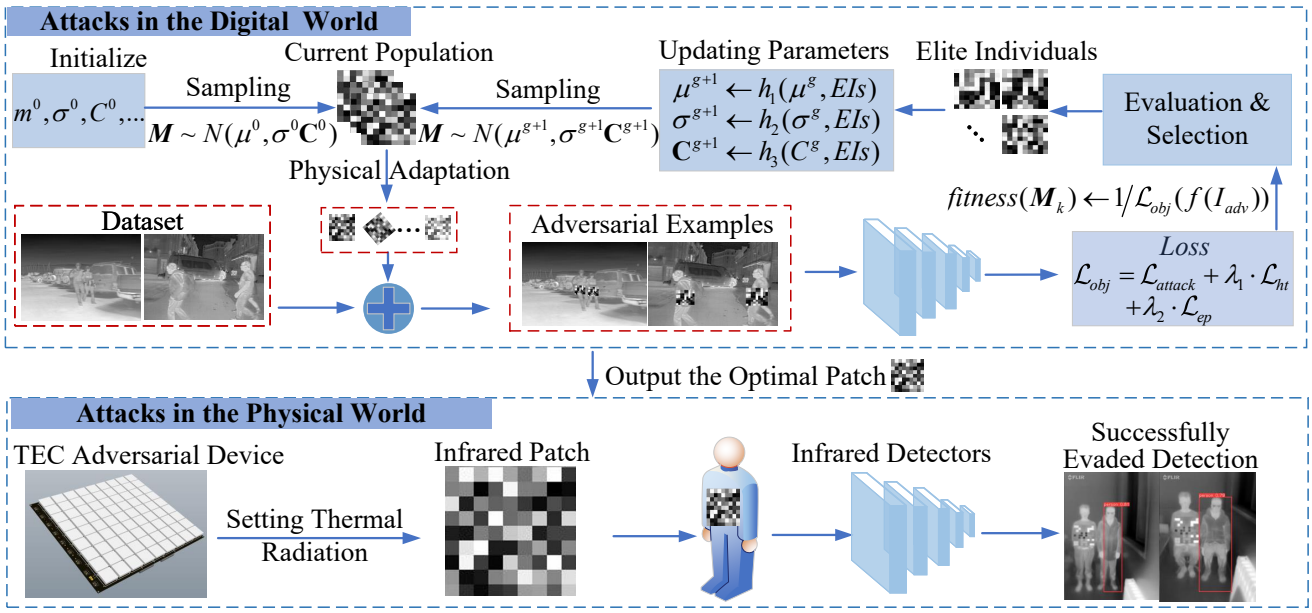

Figure 3: Overview of the proposed method. Top: attack in the digital world based on CMA-ES optimization. Bottom: attack in the physical world based on our *AdvDisplay* device.

$Q_{convection}$ is caused by the flow of a fluid, and according to Newton's Law of Cooling, it can be expressed as

$$Q_{convection} = h \times A \times |T_{obj} - T_{temp}|, \tag{4}$$

where h and A represent the convective heat transfer coefficient and the surface area in contact with the fluid, respectively. $Q_{conduction}$ represents the heat loss due to heat transfer within the object, according to Fourier's law (Casati and Li 2006), which can be written as

$$Q_{conduction} = k \times A \times \frac{\Delta T}{d}, \tag{5}$$

where k and d represent the heat conductivity and the thickness of the TEC, respectively. ΔT represents the temperature difference between the two sides of the TEC. Consider the case of well dissipated heat, $\Delta T = |T_{obj} - T_{temp}|$. Since both $Q_{convection}$ and $Q_{conduction}$ are first-order linearly related to the difference between T_{obj} and T_{temp}, the total electric power loss $\mathcal{L}_{ep}$ can be expressed as

$$\mathcal{L}_{ep} = \sum_{i,j} |T_{obj}^{i,j} - T_{temp}|. \tag{6}$$

Combine the three losses mentioned above as our total loss and balance the losses with weighted factors λ_1 and λ_2. The total loss can be written as

$$\mathcal{L}_{total} = \mathcal{L}_{obj}((f(I_{adv}, \theta)) + \lambda_1 \mathcal{L}_{ht} + \lambda_2 \mathcal{L}_{ep}. \tag{7}$$

Optimization The focus of the proposed attack is to search for the vector $\boldsymbol{M}$ of radiation intensities of the TEC array under a given target, aiming to evade detection by the detector. In this paper, the most practical scenario is considered: the attacker does not have access to the target model's knowledge and can only obtain the confidence scores of detected targets through querying the object detector. Therefore, employing popular gradient descent optimization algorithms to solve this black-box attack problem is impractical. Inspired by the work of Dong *et al.* (Dong et al. 2019), the covariance matrix adaptation evolution strategy (CMA-ES) is employed to search for the vector $\boldsymbol{M}$ to generate adversarial TEC arrays. Overall, the adversarial attack is implemented by solving the following objective:

$$\begin{aligned} \boldsymbol{M}^* &= arg \min_{\boldsymbol{M}} \mathcal{L}(f(x_{adv})), \\ s.t.\ \boldsymbol{M} &\in [\epsilon_{min}, \epsilon_{min}], \end{aligned} \tag{8}$$

where ϵ_{min} and ϵ_{max} represent the lower and upper bounds of the temperature achievable by the TEC unit, respectively. Specifically, a population is first initialized in the search space, where each individual represents a candidate $\boldsymbol{M}$. Subsequently, the covariance matrix $\boldsymbol{C}$ is introduced from a multidimensional normal distribution to guide the evolutionary direction of the population. The update of the population can be expressed as follows:

$$\boldsymbol{M}_k^{g+1} = \boldsymbol{\mu}^g + \boldsymbol{\sigma}^g N(\mathbf{0}, \boldsymbol{C}^g), \quad k = 1, ..., \lambda, \tag{9}$$

where $\boldsymbol{M}_k^{g+1}$ represents the kth individual of the $g+1$th generation of the population. $\boldsymbol{\mu}^g$ and $\boldsymbol{\sigma}^g$ denote the mean of the elite individuals of the gth generation and the global step size of the gth generation, respectively. $\boldsymbol{C}^g$ is defined as the covariance of the distribution of the gth generation of the population. λ represents the size of the population. The overall optimization process is shown in **Algorithm 1**. CMA-ES is able to model the local geometry of the search direction, especially in non-separable search spaces, and it can accelerate the optimization process by adjusting multiple variables simultaneously.

Precise Temperature Control

The proposed physical adversarial board implements precise temperature control based on the Proportional-Integral-Derivative (PID) control algorithm. The PID algorithm is a

Algorithm 1: Optimization for *AdvDisplay*.

Input: Dataset D, Detector f.
Output: The optimal $\boldsymbol{M}^*$ representing *AdvDisplay* .

Initialize relevant parameters such as $\boldsymbol{\mu}^0$, $\boldsymbol{\sigma}^0$ and $\boldsymbol{C}^0$.
Let $g = 0$
while *termination criterion is not satisfied* **do**
  Sample $\boldsymbol{P} \sim \mathcal{N}(\boldsymbol{\mu}^g, \boldsymbol{\sigma}^g \boldsymbol{C}^g)$
  while $\boldsymbol{M}_k = iterator(\boldsymbol{P})$ *is not Null* **do**
    for each I in D **do**
      $I_{adv} \leftarrow apply(I, \boldsymbol{M}_k)$
    end
    $fitness(\boldsymbol{M}_k) \leftarrow 1/\mathcal{L}_{total}(f(I_{adv}, \theta))$
  end
  $\{\boldsymbol{\mu}^{g+1}, \boldsymbol{\sigma}^{g+1}, \boldsymbol{C}^{g+1}\} \leftarrow$ update $\{\boldsymbol{\mu}^g, \boldsymbol{\sigma}^g, \boldsymbol{C}^g\}$
  Let $g = g + 1$
end
$\boldsymbol{M}^* = \max_{\boldsymbol{M}_k} fitness(\boldsymbol{M}_k)$
return $\boldsymbol{M}^*$

widely used method in industrial control systems, and its fundamental principle involves calculating the controller's output based on the error between the current state of the controlled object and the desired state. The control principle of PID is commonly expressed as

$$u(t) = K_p \cdot e(t) + K_i \cdot \int_0^t e(\tau)\, d\tau + K_d \cdot \frac{d}{dt} e(t), \quad (10)$$

where $u(t)$ is the controller's output, $e(t)$ is the control error, which is the difference between the set temperature and the current temperature. K_p, K_i, and K_d are the proportional, integral, and derivative gains of the PID controller, respectively. The proportional term $K_p \cdot e(t)$ addresses the magnitude of the error, the integral term $K_i \cdot \int_0^t e(\tau)\, d\tau$ eliminates the steady-state error, and the derivative term $K_d \cdot \frac{d}{dt} e(t)$ predicts future errors, thereby improving the transient response of the system.

To implement PID control using microprocessors such as Field-Programmable Gate Array (FPGA), it is necessary to convert the PID control method, represented by equation (10), into a discretized form that can be programmatically implemented. The discretized form of the PID control can be expressed as

$$u(t) = K_p e(t) + K_i \sum_{\tau=0}^{t} e(\tau) + K_d(e(t) - e(t-1)). \quad (11)$$

To further optimize the control effect and reduce the computational burden on the microprocessor, an accumulative incremental approach is adopted for outputting the control signal. The calculation of the increment can be expressed as

$$\Delta u(t) = K_p \Delta e(t) + K_i e(t) + K_d(\Delta e(t) - \Delta e(t-1)), \quad (12)$$

where $\Delta e(t) = e(t) - e(t-1)$. Subsequently, the controller's output is updated to $u(t) = u(t-1) + \Delta u(t)$. By employing this method, the control quantity can be calculated using only the error values at the moment t and its immediate preceding and succeeding moments. This approach enables precise and stable control of the TEC temperature.

In the implementation of physical adversarial attacks using TEC arrays, a challenge arises from the thermal transfer between adjacent TEC units. Despite efforts in the design phase to minimize temperature discrepancies between neighboring TECs by introducing heat transfer loss, this approach only partially resolves the issue. Therefore, an additional measure is employed during the physical implementation: applying insulating material between adjacent TECs. This strategy can further mitigate thermal interference between neighboring TECs, thereby enhancing the stability and robustness of the physical attack board.

Experiments

In this section, a comprehensive empirical study of the digital and physical worlds is conducted to verify the superiority of the proposed method.

Experimental Setup

Dataset In this experiment, the FLIR_ADAS_V2 dataset is employed to validate the effectiveness of our proposed method. Released by Teledyne FLIR, this dataset provides thermal imaging data for training and validating object detection networks. These thermal images are manually annotated and include various object types such as people, bicycles, cars, and dogs. However, given that our study focuses on people, images containing pedestrians are selected from the original dataset. Furthermore, only those images where the height of the pedestrians is more than 120 pixels are retained to ensure that the pedestrians have sufficient visibility. Finally, 1255 images are available and 878 of them are used for training and 377 images are used for testing. Then, the test images that can be successfully detected by the object detector with high probability are selected as the final attack images. Thus, the clean Average Precision (AP) is 100%. In the experiments to validate physical adversarial attacks, a FLIR ONE Pro infrared camera with the thermal resolution of 160×120 is used to capture thermal infrared images.

Detector YOLOv5 is chosen for our attack because it is a fast, effective and widely used detector[2]. We fine-tune the model based on the official pre-training weights on the filtered FLIR_ADAS_V2 dataset. The AP of the fine-tuned model reaches 99.2% on the training set and 94.5% on the testing set.

Evaluation Metrics: The proposed *AdvDisplay* aims to enable pedestrians to evade detection by infrared detectors. Therefore, AP and Attack Success Rate (ASR) are used to evaluate the attack performance of the proposed method. AP is calculated by measuring the area under the Precision-Recall (PR) curve, with a lower AP score indicating stronger attack performance. The calculation of ASR is as

$$ASR(D) = 1 - \frac{1}{N} \sum_{i=0}^{N} sign(label_i), \quad (13)$$

[2]https://github.com/ultralytics/yolov5.

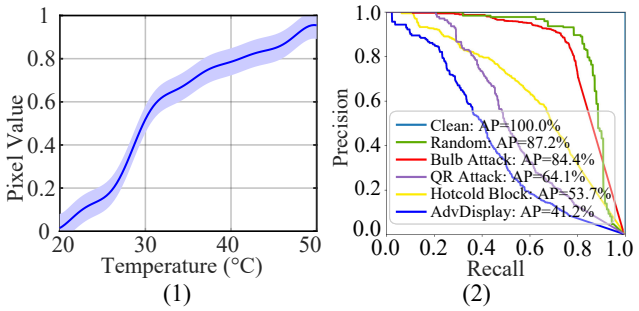

Figure 4: Quantitative results. (1) Temperature to pixel value mapping curve. (2) Precision-Recall curve in the digital space.

$$sign(label_i) = \begin{cases} 1 & label_i \in L_{pre}, \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where N represents the number of all true positive labels detected by the detector in the clean dataset, and L_{pre} is the set of all labels detected under the attack.

Other Details The TEC unit used can control temperature accurately between 20 and 50 °C. To establish the mapping of temperature in the physical world to pixel values in the digital world, 1000 images were taken by gradually increasing the temperature of the TEC unit from 20 °C to 50 °C, and then the pixel values in the TEC region of each image were averaged to construct a temperature-pixel data sample. Finally, 1000 sets of uniformly distributed data points were obtained, and then the model was constructed using Gaussian process regression. The fitting curve is shown in Figure 4(1). The fitting was good and the Root Mean Squard Error (RMSE) is 0.028.

Simulation of Physical Attacks

Comparisons with SOTA Methods To validate the effectiveness of our proposed method, we conducted control experiments using random display. Furthermore, to thoroughly analyze the attack performance of the proposed method, it is compared with three recently proposed infrared adversarial attack methods. For clarity, a brief introduction to the comparative methods is first provided. The HOTCOLD Block is an innovative physical attack method for infrared detectors, utilizing wearable heating and cooling pastes to hide persons. The Bulb Attack is a physical attack method that deceives infrared pedestrian detectors using small light bulbs. The QR Attack method is based on aerogel material and designs adversarial "QR code" patterns for multi-angle physical attacks on infrared detectors. Figure 4(2) displays the Precision-Recall (PR) curves and lists the corresponding AP values. It is observed that our method reduces the AP of pedestrian detectors by 41.2%, significantly outperforming the 87.2% of random display, thereby confirming the effectiveness of our proposed method. Additionally, our method also demonstrates competitive performance compared to currently proposed infrared adversarial attack methods, with its attack effectiveness surpassing 84.4% of the bulb attack and 64.1% of the QR attack.

Resolution	Method	s = 1cm		s = 2cm		s = 3cm		s = 4cm	
		AP	ASR	AP	ASR	AP	ASR	AP	ASR
3×3	Random	95.2	2.8	94.3	6.5	93.3	13.4	90.3	16.1
	AdvDisplay	81.9	12.2	73.2	18.7	67.2	31.0	53.3	35.1
6×6	Random	95.3	2.2	93.4	8.1	92.1	8.9	90.1	11.7
	AdvDisplay	75.2	18.7	70.3	21.9	61.4	32.1	49.9	37.2
9×9	Random	91.0	12.1	86.6	19.3	82.1	27.3	87.2	29.6
	AdvDisplay	69.6	28.9	50.1	39.7	45.1	47.3	41.2	48.2
12×12	Random	92.2	6.7	92.1	7.1	90.2	10.3	89.2	13.7
	AdvDisplay	57.2	37.1	49.3	39.8	43.2	52.7	31.1	59.6

Table 2: The attack results on the clean test set for different resolutions of *AdvDisplay* and different sizes of employed TECs.

Loss functions	AP	ASR	Heat Transfer (Normalize)	Power (W)
$\mathcal{L}_{obj}$	35.3	59.6	0.412	253.8
$\mathcal{L}_{obj} + \mathcal{L}_{ht}$	39.1	56.7	0.187	267.6
$\mathcal{L}_{obj} + \mathcal{L}_{ht} + \mathcal{L}_{ep}$	41.2	48.2	0.221	196.2

Table 3: Ablation study for different loss functions.

Ablation Study Here, the impact of different *AdvDisplay* resolutions r and TEC sizes s on the attack performance is tested. Table 2 shows the attack results of our method and random display on the clean dataset. The results show that *AdvDisplay* significantly outperforms the random display attack, confirming the effectiveness of the proposed method. Additionally, the attack ability of *AdvDisplay* improves with increased resolution and TEC size, aligning with our expectations. However, it is worth noting that excessively high resolution and excessively large adversarial patches may cause challenges for physical implementation. *AdvDisplay* with $r = 9 \times 9$ and $s = 4cm \times 4cm$ achieved 48.2% ASR and reduced AP to 41.2%. Given the effectiveness of the *AdvDisplay* attack and the cost of the physical implementation, the patch with $r = 9 \times 9$ and $s = 4cm \times 4cm$ is chosen as the primary focus of this study.

To evaluate the impact of different loss functions, we progressively incorporate $\mathcal{L}_{obj}$, $\mathcal{L}_{ht}$, and $\mathcal{L}_{ep}$ into the total loss as $\mathcal{L}_{total}$ to optimize the *AdvDisplay* . Table 3 presents the quantitative results of different loss functions. It is noted that by combining $\mathcal{L}_{ht}$ and $\mathcal{L}_{ep}$, the ASR slightly decreases from 59.6% to 48.2% but shows significant improvement in decreasing heat transfer and reducing power. This implies that the designed joint loss can achieve better physical implementability and lower adversarial costs without significantly impacting the attack performance.

Attacks in the Physical World

To thoroughly test the adversarial performance of *AdvDisplay* in the physical world, our experimental design incorporated variations in distance, angle, posture, and scene, with the AP and ASR being calculated through recorded videos. By default, we let a volunteer stand with *AdvDisplay* at a distance of 2m from the infrared camera for the test. Additionally, under the same conditions, another volunteer served as a control group without holding anything. Figure 5 presents examples of some test results. It can be observed that the volunteer holding our

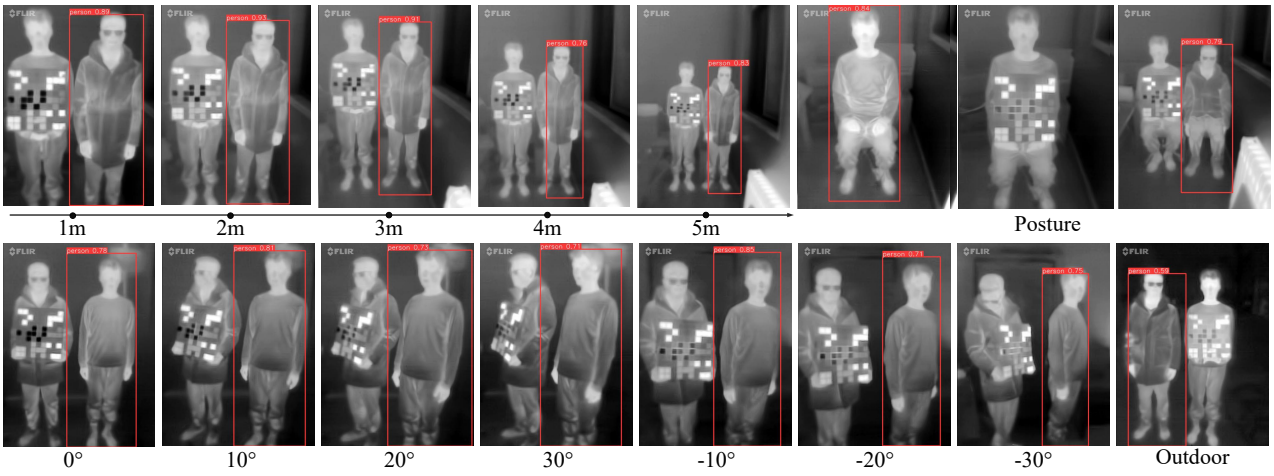

Figure 5: The design and production of AdvDisplay device.

designed *AdvDisplay* can successfully evade the infrared detector. Meanwhile, the volunteer holding nothing cannot achieve invisibility in front of the detector. Statistically, our *AdvDisplay* achieves 96.1% ASR in the default setting. The experimental results show that the designed *AdvDisplay* can successfully attack real-world infrared detection systems.

To further verify the robustness of *AdvDisplay* , we conducted more comprehensive tests by changing distance, angle, posture, and scene based on the default conditions. For the shooting distance, the volunteer gradually moved from 1m to 5m away from the infrared camera. For the shooting angle, the volunteer gradually rotated from $\pm 0°$ to $\pm 30°$. For the shooting posture, the volunteer changed from standing to squatting. For shooting scenes, we chose both indoor and outdoor settings for testing. Figure 5 shows visual examples of these conditions. Figure 6 gives the results of AP and ASR at different distances and angles. It can be found that when the shooting distance changed from 1m to 5m, the ASR of *AdvDisplay* decreased by 54.5%. The decrease in ASR is relatively significant as the shooting distance increases. Observing the given visualized examples of physical adversarial attacks, it can be noticed that this result is due to the lack of resolution of the employed infrared camera, which causes the details of *AdvDisplay* to become blurred. Nevertheless, the ASR is still around 42%. When the shooting angle changed, the ASR of *AdvDisplay* remained high. In addition, when the shooting scene changed from indoor to outdoor, the ASR of *AdvDisplay* only decreased by 12.1%. These results indicate that *AdvDisplay* exhibits strong robustness in different scenarios.

Adversarial Defense Test

Here, the method for defending against *AdvDisplay* is tested. Adversarial training, a widely adopted and effective defense technique, aims to enhance the model's resilience to adversarial attacks (Bai et al. 2021). We employ this method to defend against *AdvDisplay* attacks. Specifically, adversarial samples generated based on *AdvDisplay* are merged into the clean dataset, and the YOLOv5 model is fine-tuned accordingly. After adversarial training, the model's AP on the adversarial samples increases from 41.2% to 73.6%, and the ASR decreases from 48.2% to 29.7%. Adversarial training enhances the model's robustness, indicating that our work can be used to improve model performance further. However, adversarial training only reduces the threat posed by *AdvDisplay* to a certain extent, which also demonstrates the robustness of our method.

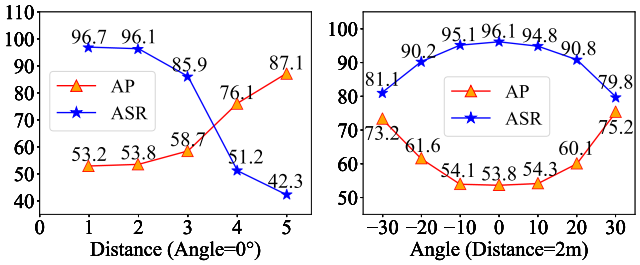

Figure 6: Analysis of attacks at different distances and angles.

Conclusion

In this paper, the TEC material is applied for the first time in infrared physical adversarial attacks. With the design of the temperature control algorithm and circuitry, a reusable physical adversarial device is fabricated in which radiation intensity can be precisely, continuously, and bidirectionally adjusted. Compared to current methods, *AdvDisplay* possesses a more extensive search space and can generate more adversarial patterns. In addition, heat transfer loss and electric power loss are proposed to decrease the difference between adversarial patches in the digital and physical worlds and the power of *AdvDisplay* devices. Extensive empirical studies in both the digital and physical worlds show that our *AdvDisplay* possesses superior adversarial effectiveness compared to the existing methods and has strong robustness.

Acknowledgments

This project is supported by National Natural Science Foundation of China No. 62036006 and 62306221.

References

Bai, T.; Luo, J.; Zhao, J.; Wen, B.; and Wang, Q. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. *arXiv preprint arXiv:2102.01356*.

Bendelac, S.; Manville, K.; Harguess, J.; and Rodriguez, M. 2021. A Dynamic Thermal IR Display for Physical Adversarial Attacks. In *Proc. Appl. Imagery Pattern. Recogn. Workshop*, 1–7.

Bustos, N.; Mashhadi, M.; Lai-Yuen, S. K.; Sarkar, S.; and Das, T. K. 2023. A systematic literature review on object detection using near infrared and thermal images. *Neurocomputing*, 560: 126804.

Casati, G.; and Li, B. 2006. Heat conduction in one dimensional systems: Fourier law, chaos, and heat control. In *Proc. Non-Linear Dynamics and Fundamental Interactions*, 1–16.

Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; and Zhu, J. 2019. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. In *Proc. CVPR*.

Gong, X.; Yuan, D.; and Bao, W. 2021. Understanding Partial Multi-Label Learning via Mutual Information. In *Proc. Adv. neural inf. proces. syst.*, volume 34, 4147–4156.

Gong, X.; Yuan, D.; and Bao, W. 2023. Discriminative Metric Learning for Partial Label Learning. *IEEE Trans. Neural Networks Learn. Sys.*, 34(8): 4428–4439.

Gong, X.; Yuan, D.; Bao, W.; and Luo, F. 2023. A Unifying Probabilistic Framework for Partially Labeled Data Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7): 8036–8048.

Hu, C.; Shi, W.; Jiang, T.; Yao, W.; Tian, L.; Chen, X.; Zhou, J.; and Li, W. 2024. Adversarial infrared blocks: A multi-view black-box attack to thermal infrared detectors in physical world. *Neural Netw.*, 175: 106310.

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 *Proc. CVPR*, 13307–13316.

Huang, Y.; Dong, Y.; Ruan, S.; Yang, X.; Su, H.; and Wei, X. 2024. Towards Transferable Targeted 3D Adversarial Attack in the Physical World. In *Proc. CVPR*, 24512–24522.

Kim, T.; Lee, H. J.; and Ro, Y. M. 2022. Map: Multispectral Adversarial Patch to Attack Person Detection. In *Proc. ICASSP*, 4853–4857.

Kim, T.; Yu, Y.; and Ro, Y. M. 2023. Multispectral Invisible Coating: Laminated Visible-Thermal Physical Attack against Multispectral Object Detectors Using Transparent Low-E Films. In *Proc. AAAI*, volume 37, 1151–1159.

Kou, R.; Wang, C.; Peng, Z.; Zhao, Z.; Chen, Y.; Han, J.; Huang, F.; Yu, Y.; and Fu, Q. 2023. Infrared small target segmentation networks: A survey. *Pattern Recogn.*, 143: 109788.

Li, Z.; Liu, F.; Yang, W.; Peng, S.; and Zhou, J. 2022. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Networks Learn. Sys.*, 33(12): 6999–7019.

Liu, W.; and Tsang, I. 2015. On the Optimality of Classifier Chain for Multi-label Classification. In *Adv. neural inf. proces. syst.*, volume 28.

Liu, W.; and Tsang, I. W. 2017. Making decision trees feasible in ultrahigh feature and label dimensions. *J. Mach. Learn. Res.*, 18(81): 1–36.

Ma, W.; Li, Y.; Jia, X.; and Xu, W. 2023. Transferable Adversarial Attack for Both Vision Transformers and Convolutional Networks via Momentum Integrated Gradients. In *Proc. ICCV*, 4630–4639.

Sidebotham, G. 2015. Heat Transfer Modes: Conduction, Convection, and Radiation. *Heat Transfer Modeling: An Inductive Approach*, 61–93.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wei, H.; Wang, Z.; Jia, X.; Zheng, Y.; Tang, H.; Satoh, S.; and Wang, Z. 2023a. HOTCOLD Block: Fooling Thermal Infrared Detectors with a Novel Wearable Design. In *Proc. AAAI*, volume 37, 15233–15241.

Wei, X.; Huang, Y.; Sun, Y.; and Yu, J. 2023b. Unified Adversarial Patch for Cross-Modal Attacks in the Physical World. In *Proc. ICCV*, 4445–4454.

Wei, X.; Yu, J.; and Huang, Y. 2023. Physically Adversarial Infrared Patches With Learnable Shapes and Locations. In *Proc. CVPR*, 12334–12342.

Wilson, A. N.; Gupta, K. A.; Koduru, B. H.; Kumar, A.; Jha, A.; and Cenkeramaddi, L. R. 2023. Recent Advances in Thermal Imaging and its Applications Using Machine Learning: A Review. *IEEE Sensors J.*, 23(4): 3395–3407.

Wu, W.; Su, Y.; Lyu, M. R.; and King, I. 2021. Improving the Transferability of Adversarial Samples With Adversarial Transformations. In *Proc. CVPR*, 9024–9033.

Xiong, Z.; Xu, H.; Li, W.; and Cai, Z. 2021. Multi-Source Adversarial Sample Attack on Autonomous Vehicles. *IEEE Trans. Veh. Technol.*, 70(3): 2822–2835.

Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; and Pi, Y. 2024. EFLNet: Enhancing Feature Learning Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–11.

Yuan, X.; He, P.; Zhu, Q.; and Li, X. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. Neural Networks Learn. Sys.*, 30(9): 2805–2824.

Zhang, X.; and Demiris, Y. 2023. Visible and Infrared Image Fusion Using Deep Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8): 10535–10554.

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 *Proc. CVPR*, 13317–13326.

Zhu, X.; Hu, Z.; Huang, S.; Li, J.; Hu, X.; and Wang, Z. 2023. Hiding from infrared detectors in real world with adversarial clothes. *Appl. Intell.*, 53(23): 29537–29555.

Zhu, X.; Li, X.; Li, J.; Wang, Z.; and Hu, X. 2021. Fooling Thermal Infrared Pedestrian Detectors in Real World Using Small Bulbs. In *Proc. AAAI*, volume 35, 3616–3624.

Zhu, X.; Li, X.; Li, J.; Wang, Z.; and Hu, X. 2024. Hiding from thermal imaging pedestrian detectors in the physical world. *Neurocomputing*, 564: 126923.