Physical Backdoor: Towards Temperature-based Backdoor Attacks in the Physical World

Wen Yin[1,3], Jian Lou[4], Pan Zhou[1*], Yulai Xie[1,2,3 *], Dan Feng[2,3],
Yuhua Sun[1], Tailai Zhang[1,3], Lichao Sun[5]
[1]School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[3] Jinyinhu Laboratory [4]Zhejiang University [5]Lehigh University
{wenyin, panzhou, ylxie, dfeng, natsun, tl_zhang}@hust.edu.cn
jian.lou@zju.edu.cn, james.lichao.sun@gmail.com

Abstract

Backdoor attacks have been well-studied in visible light object detection (VLOD) in recent years. However, VLOD can not effectively work in dark and temperature-sensitive scenarios. Instead, thermal infrared object detection (TIOD) is the most accessible and practical in such environments. In this paper, our team is the first to investigate the security vulnerabilities associated with TIOD in the context of backdoor attacks, spanning both the digital and physical realms. We introduce two novel types of backdoor attacks on TIOD, each offering unique capabilities: Object-affecting Attack and Range-affecting Attack. We conduct a comprehensive analysis of key factors influencing trigger design, which include temperature, size, material, and concealment. These factors, especially temperature, significantly impact the efficacy of backdoor attacks on TIOD. A thorough understanding of these factors will serve as a foundation for designing physical triggers and temperature controlling experiments. Our study includes extensive experiments conducted in both digital and physical environments. In the digital realm, we evaluate our approach using benchmark datasets for TIOD, achieving an Attack Success Rate (ASR) of up to 98.21%. In the physical realm, we test our approach in two real-world settings: a traffic intersection and a parking lot, using a thermal infrared camera. Here, we attain an ASR of up to 98.38%.

1. Introduction

Thermal infrared object detection (TIOD) have several unique advantages over visible light object detection (VLOD). It excels in detecting objects under low visible light, smoky, heavy rain, and intense snow environments, making it less affected by glare and light mutation, all while retaining its sensitivity to thermal changes in objects [15]. Consequently, TIOD becomes increasingly indispensable in a variety of application scenarios, from security monitoring and autonomous driving in the dark to temperature measurement during a pandemic. The security vulnerabilities of VLOD are thoroughly examined for both adversarial attacks [14, 17] and backdoor attacks [18, 34, 45]. Backdoor attacks manipulate both a small portion of contaminated training samples and testing samples with the backdoor trigger. Then, the trained detector will have backdoor effects at the testing time when encountering samples with the backdoor trigger, while remaining normal when fed with clean testing samples. In real-world scenarios, backdoor attacks pose a serious security threat to deep neural networks (DNNs) [31, 52, 53] due to their stealthiness.

Unlike VLOD field [13, 59], the security vulnerabilities of TIOD remain largely unexplored and current efforts are focused merely on adversarial attacks rather than backdoor attacks. For instance, Zhu et al. design adversarial patterns and manufacture an adversarial shirt made of aerogel material [63]. Wei et al. introduce the method of aggregation regularization to optimize the adversarial infrared patch, making the patch easier to implement physically [50]. Both works suggest that extra considerations are required to design effective adversarial examples for TIOD, mainly due to the unique characteristics of thermal infrared imaging compared to visible light imaging. These new adversarial attacks ring the alarm that TIOD demands the same level of scrutiny as VLOD to expose all types of potential security threats. However, the security vulnerabilities of TIOD to backdoor attacks remain unexplored.

TIOD combines object detection technology with thermal infrared imaging technology, allowing it to recognize

*Corresponding author.

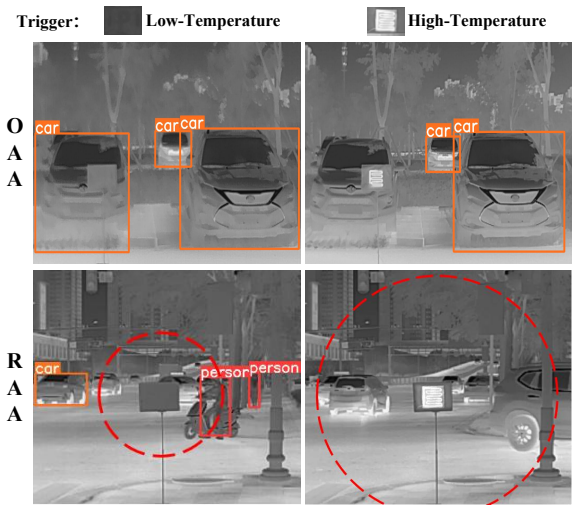

Figure 1. Examples of OAA and RAA for *car disappearance* in the physical world. By changing the temperature of the trigger, it is capable to switch between whether the backdoor is activated in OAA and to adjust the attack range in RAA. The affecting range of RAA is marked as the red circle.

objects captured using infrared thermal radiation imaging. Unlike RGB images with three channels, thermal infrared images have only a gray-scale channel and contain less texture information [62]. Backdoor attacks on VLOD can take advantage of the additional information lying in the extra channels to gain more capacity in trigger design, which ultimately allows for a strong attack capability. The design of backdoor attacks on TIOD becomes more challenging than the visible light counterpart, because the design space for the trigger is restricted to properly placing the trigger, choosing a material with ideal thermal infrared characteristics, or manipulating its temperature. Therefore, it raises the following compelling question: *Can we design effective backdoor attacks on TIOD by utilizing their unique properties compared to VLOD?*

In this paper, we propose two backdoor attack methods against TIOD : Object-Affecting Attack (OAA) and Range-Affecting Attack (RAA). OAA manipulates the detection results for a specific object carrying with a trigger. While RAA causes all objects of a chosen class in close proximity to the trigger to be misidentified. In addition, we propose a new mechanism to adjust the triggering behavior of backdoor attacks by temperature modulating. The demonstration result is presented in Figure 1. Remote control of backdoor attacks involves a simple button press, eliminating the need for any visible light visual changes to the prearranged scene. Since the temperature change of the object is not visible to the human eye, temperature modulating offers more stealthiness and flexibility. Our in-depth study of proposed attacks in the digital world, and successful implementation in the physical world, provides an affirmative answer to the question in the preceding paragraph that dedicated backdoor attacks indeed pose significant threats to TIOD. The main contributions of this paper are as follows:

- We examine the security vulnerability of TIOD to backdoor attacks and identify the critical factors that differentiate their trigger design from that of VLOD. To the best of our knowledge, this is the first study of backdoor attacks on TIOD.
- We propose two types of backdoor attacks of OAA and RAA that offer different affecting capacities. In addition, we further propose a novel backdoor trigger by modulating its temperature, allowing the backdoor effect to be activated or deactivated within different temperature ranges in OAA and adjusting the affecting range in RAA.
- In a digital environment, we validate the attack's effectiveness across various parameters, achieving an ASR of up to 98.21%. In the physical world, we test the proposed backdoor attacks in two representative real scenes of a traffic intersection and a parking lot. Our attacks are effective in both scenes, achieving an average ASR of over 96%. In addition, the methods are cost-friendly, with the production of an electric heating device as a trigger costing less than 5 US dollars. We also evaluate three potential countermeasures defending against our attacks.

2. Related Work

Thermal Infrared Object Detection. Object detection is a fundamental task in computer vision, which also serves as the foundation of image segmentation [6], object tracking [36], and keypoint detection [2]. TIOD uses image information in the thermal infrared domain for object recognition [21]. The YOLO model [39], as the first one-stage detector, is employed for pedestrian detection in the thermal infrared domain [25]. The classic two-stage detector, Faster RCNN, is also applied for TIOD, although its detection accuracy is limited [11]. In diverse environments, including severe weather conditions, Huda et al. achieve accurate pedestrian detection using YOLO v3 and thermal infrared data [23]. Currently, YOLO v3 is integrated into the latest versions of autonomous driving systems such as Apollo [3]. Gong et al. conduct a study on vehicle recognition in the thermal infrared domain to assist autonomous driving systems [16]. Furthermore, Wang et al. combine YOLO v5 to enhance the detection accuracy of TIOD [48]. There is a growing body of research that focuses on the performance improvement and data diversity in TIOD [10, 26].

Backdoor Attack. Backdoor attacks are carried out by embedding hidden backdoors into DNNs The attacker generates a backdoor model through "data poisoning" attacks [1, 22, 35, 57] or "poisoning + training manipulation" attacks [8, 37, 55] (the former only poisons the training data, while the latter not only poisons the training data but also modifies the training process).The triggers used to activate backdoor effect are also diverse. Existing triggers of back-

door attacks include single pixel [46], reflection background [35], invisible patterns [5, 29, 41], and so on. In addition to directly poisoning training data, backdoor attacks can also embed hidden backdoors by modifying model weights through transfer learning [28, 49]. Therefore, backdoored model training can occur in all steps of the training process, which is a serious threat to DNNs [19, 42, 43, 56, 58]. The research into backdoor attacks drive the development of defense methods [20, 44]. Beyond digital world, there are also works exploring backdoor attacks in the physical world [7, 18, 51, 61]. The above works all focus on the visible light field, while backdoor attacks in the thermal infrared field still lack exploration.

3. Backdoor Attacks on TIOD

3.1. Threat Model

Attacker's Goal. Our backdoor attack has two goals, both of which align with common design principles of backdoor attacks. The first goal is for stealthiness purpose to ensure that the backdoored TIOD can still properly identify the objects in clean samples. This task can make it difficult for the user to find the model anomalies without knowing the backdoor trigger. The second goal is for effectiveness purpose to cause the backdoored TIOD to either not identify the object (i.e., *object disappearance*) or identify it with an incorrect object class (i.e., *object misclassification*) in backdoor samples with the attacker-chosen trigger. In this paper, we focus on attacking cars as the exemplary object class, due to the serious security consequences in real-world applications such as autonomous driving.

Attacker's Capability. Backdoor attacks can occur in many situations, such as outsourcing training, using pretrained models for transfer learning, and collecting data from unknown sources. Following previous work on backdoor attacks [9, 32], we adopt the "data poisoning" threat model. It suffices to gain access to part of the training dataset in order to inject poisoned training samples, while leaving the training process untempered. Since this threat model does not interfere with the model and training process, it can fundamentally reveal the vulnerability of TIOD, thereby promoting research on the improvement of object detection security, such as backdoor defense methods.

3.2. Problem Formulation

In this paper, we mainly attack the object detection model YOLO v5. The YOLO v5 contains three loss components as follow: classification loss L_{cls}, BBox regression loss L_B [60], and confidence loss L_{conf}. The classification loss is calculated only when there is an object in the detection BBox. Please refer to Appendix A for the detailed formulation of each loss component. The total loss function is composed of all three individual loss components,

$$L = \alpha L_{cls} + \beta L_{conf} + \gamma L_B, \tag{1}$$

where α, β, and γ are the balancing hyper-parameters.

Backdoor attackers can poison a small portion q of the training dataset. Without loss of generality, let the poisoned training dataset $\widetilde{S}$ be divided into a clean dataset S_c and a dirty dataset S_d, where $|S_d| = q * |\widetilde{S}|$ with $|\cdot|$ denoting the cardinality of the datasets. The poisoned images in the dirty dataset are modified from the original images with the attacker-chosen trigger injected, denoted by x_i^d. The dirty label l_i^d is obtained by modifying the original label l_i according to the attack purpose (i.e., misclassification or disappearance), which is expressed as follows :

$$l_i^d = \begin{cases} l_{oc} & \textit{Object Misclassification} \\ None & \textit{Object Disappearance}, \end{cases} \tag{2}$$

where l_{oc} indicates that the class of the label is replaced, and $None$ indicates that the label is deleted. Altogether, the dirty training sample in the dirty dataset becomes (x_i^d, l_i^d).

During training, neuron activations in the network will become abnormally affected, causing the input image with the trigger to be incorrectly mapped to a specific output. Depending on the attack purpose, the backdoor attack can be formulated by one of the two optimization problems. For *Object Misclassification*,

$$\underset{\widetilde{\theta}}{argmin}\, L = \underset{\widetilde{\theta}}{argmin}(\alpha L_{cls}(\widetilde{S}, \widetilde{\theta}) + \beta L_{conf} + \gamma L_B), \tag{3}$$

where $L_{cls}(\widetilde{S}, \widetilde{\theta})$ represents the classification loss on the poisoned training set $\widetilde{S}$ with backdoored model parameter $\widetilde{\theta}$. Since only the classification label of the object is modified, the poisoned training dataset has no direct relationship with the confidence loss and BBox regression loss. For *Object Disappearance*,

$$\underset{\widetilde{\theta}}{argmin}\, L = \underset{\widetilde{\theta}}{argmin}(\alpha L_{cls} + \beta L_{conf}(\widetilde{S}, \widetilde{\theta}) + \gamma L_B), \tag{4}$$

where $L_{conf}(\widetilde{S}, \widetilde{\theta})$ represents the confidence loss on the poisoned dataset $\widetilde{S}$ with backdoored model parameters $\widetilde{\theta}$. Recall that the classification loss and BBox regression loss are involved only when the confidence loss indicates a high level of object existence confidence. Therefore, by removing the object class label, we can disrupt the normal function of the confidence loss component at the presence of the trigger, causing the detector to falsely identify the object as disappearance in detection.

3.3. Proposed Attacks

Overview. The attack procedure is illustrated in Figure 2. Both attack methods are implemented by poisoning a subset of the training set with the following two steps: 1) Trigger Insertion and 2) Label Modification. The backdoor model trained by such a poisoned dataset can achieve the attack effect required by the attacker.

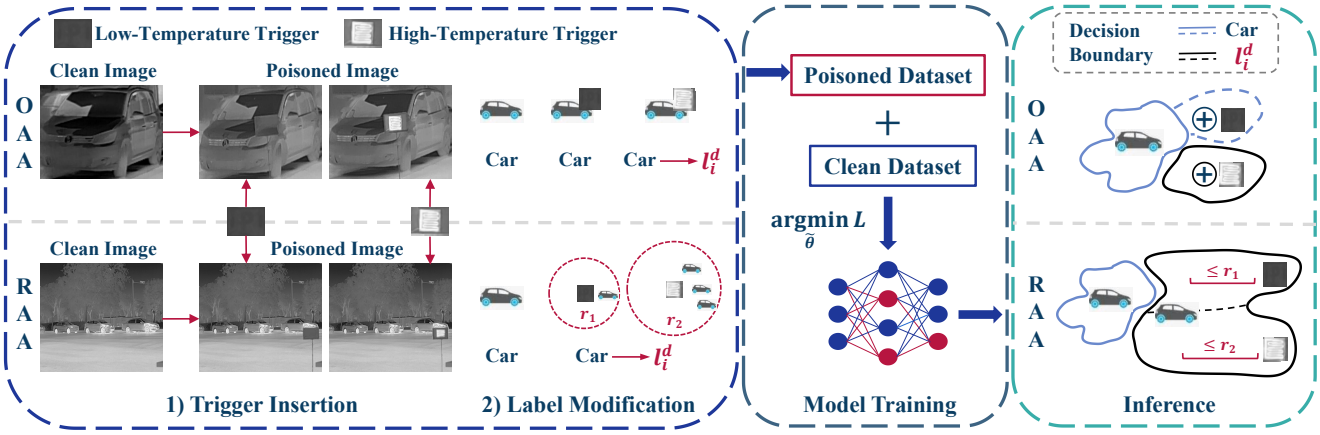

Figure 2. Overview of our proposed attacks. The red arrow in 2) indicates that "car" is modified to "l_i^d". The r_1 and r_2 are attack radius.

3.3.1 Preliminaries

Temperature-Pixel Value Mapping. Thermal infrared cameras with different operating wavelengths have different response functions. Empirically, we obtain the approximate function as follows [24],

$$p = g(t) = \Lambda T^m + \Phi, \tag{5}$$

where $p \in [0, 255]$ is the pixel value, T is the object temperature, and p increases as T increases. Λ and Φ are the adjustment parameters obtained by the thermal infrared camera. The operating wavelength in this paper is between 8 and 13 μm, so $m = 3.9889$ [54]. In practical applications, $m = 4$ has little effect on the measurement results. As

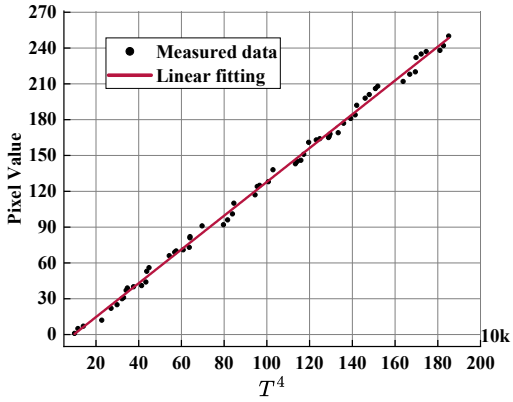

Figure 3. Function fitting of $p - T^4$.

shown in Figure 3, we measure temperatures corresponding to different pixel values across multiple thermal infrared images. Subsequently, we perform linear fitting on the measured data to establish the following mapping relationship,

$$p = 1.4221 * 10^{-4} * T^4 - 15.4760. \tag{6}$$

Due to the correction function of the thermal infrared camera, the temperature of the same point measured with a thermal infrared camera does not vary with distance [62].

Trigger Design. Existing backdoor attacks for VLOD rely on triggers designed based on color differences [4]. When applied to the thermal infrared domain, as shown in Figure 4, such triggers will appear as grayscale patterns. Consequently, their intricate texture details cannot be effectively kept, rendering backdoor attacks unable to be effectively triggered. To overcome this limitation, the designed trigger needs to have a temperature attribute difference from the attacked target, and the design of backdoor attacks should be based on the temperature difference. After comparing insulation cotton sheets, plastic sheets, and electric heaters, we find that electric heaters offer better control over temperature changes. As shown in Figure 5, the physical trigger is an electric heating device consisting of an electric heater and a signboard. The choice of sign can be adjusted to suit the scene and ensure unobtrusiveness. The temperature can be remotely controlled with a single button press. The morphology of triggers at different temperatures in the thermal infrared world is shown in Figure 6. Given the uniform heat distribution of the object, we can simulate digital triggers using pixel blocks based on Equation (6). This approach facilitates the exploration of parameter influences on attacks.

Figure 4. RGB triggers mapped to the thermal infrared domain.

3.3.2 Object-Affecting Attack

How to ensure the effectiveness of the attacks and maintain the Benign Accuracy (BA) [30] of the model are key issues for backdoor attacks. Since the trigger should be added inside the object BBox which can be small (resp. large) for object lying remote (resp. close) to the infrared camera, we adjust the trigger size according to this object's BBox in

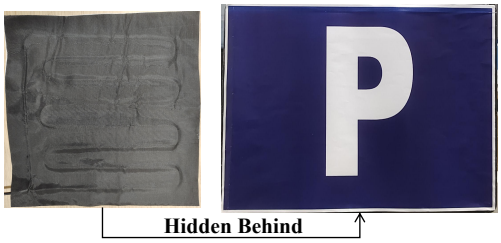

Figure 5. Physical electrothermal trigger design.

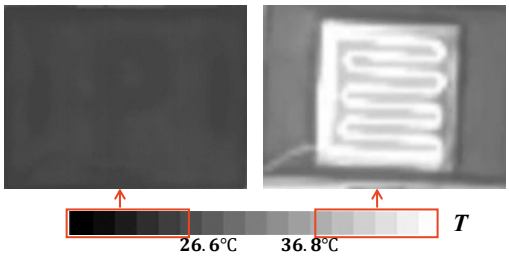

Figure 6. The trigger in the thermal infrared camera.

order to simulate the visual effect of the trigger in the physical world. Given the object o and the size of its BBox s, the object after trigger insertion is

$$o' = o + y(p, \lambda s), \tag{7}$$

where trigger $y(p, \lambda s)$ is a pixel block with pixel value p and size λs. Then, we attach triggers to all objects from the targeted class (e.g., all cars in the image), which can be expressed as follows:

$$x' = x + \sum_{class=car} y(p, \lambda s), \tag{8}$$

where x' is the image after adding all triggers to clean image x. Assuming the attack pixel range is $[p_1, p_2]$, if $p \in [p_1, p_2]$, we perform Label Modification to this object. As shown in Equation (2), according to different attack purpose, we can either delete the label or replace the class of the label to generate the poisoned dataset required for training the backdoor model.

3.3.3 Range-Affecting Attack

The one-stage detectors unusually divide the image into many grids during detection. Whenever the object under detection has some overlap with any grids, these grids will all participate in the detection of this object. Since an object may occupy multiple adjacent grids, it is possible to add the trigger close to the object so that the grids occupied by the trigger will overlap with some of the object's grids, even though the trigger does not strictly lie inside the object's BBox. Therefore, backdoor attack can still establish the abnormal association between the trigger and the object *indirectly* via the overlapping grids, instead of *directly* putting inside the object's BBox. Based on this analysis, we propose the second backdoor attack called RAA.

For the poisoned dataset, we insert only a single trigger pattern into each poisoned image. The location of the trigger in RAA can be more flexible. By observing objects with our thermal infrared camera utilized in the physical experiments, we select common telephone poles or street signs on the road as triggers, which can be simulated with pixel strips and pixel blocks in the digital world, respectively. Given a pixel strip $y(p, hw)$ with length h, width w and pixel value $p \in [0, 255]$ as a trigger, the attacker arbitrarily selects a point (a, b) in the clean image x as the center point of the trigger to insert the trigger into the clean image. Finally, the poisoned image can be obtained as follows,

$$x' = x + y(p, hw, (a, b)). \tag{9}$$

Different from OAA, RAA needs to modify the labels of all object within a certain range from the trigger. Given the following correspondence between the pixel value and the attack radius,

$$(p, ar) = \{(p_1, r_1), (p_2, r_2), ..., (p_n, r_n)\}. \tag{10}$$

If $p = p_n$ and the object's coordinate (a_o, b_o) satisfies the following conditions,

$$\sqrt{(a_o - a)^2 + (b_o - b)^2} \leq r_n, \tag{11}$$

we perform Label Modification on the object according to Equation (2). Then, we get the poisoned dataset and use it as training input to get the backdoor model of RAA. RAA allows an attacker adjustable attack affecting range, which makes the attack more flexible.

4. Experiments

In this section, we conduct experimental evaluations of OAA and RAA in both digital and physical worlds.

Evaluation Metrics. Since there is no existing intuitive effectiveness metrics for the backdoor object detection attack, we introduce a new metric - Benign Accuracy Fluctuation (BAF), in addition to the commonly adopted metric of Attack Success Rate (ASR). BAF is the value obtained by subtracting the mAP of clean samples tested by the clean model from that returned by the backdoor model. We add trigger to the objects in the test image regardless of the object size, and then define the trigger addition as successful when the following conditions are met.

$$\begin{aligned} &IOU(GT, BS) \geq IOU_{model}, \\ &Class(GT) = Class(BS) = car, \end{aligned} \tag{12}$$

where GT is the ground truth label of the object, and BS is the label obtained by the backdoor model detecting clean samples. IOU_{model} is the IOU threshold set during model detection. The total number of successful trigger additions is recorded as $\sum N_{ta}$. For the objects successfully added trigger, we define the attack as successful when they meet the following conditions at the same time. When the attack setting is to misidentify the car as the person,

$$IOU(GT, DS) \geq IOU_{model}, Class(DS) = person. \tag{13}$$

When the attack setting is that the car cannot be detected,

$$IOU(GT, DS) < IOU_{model}, \tag{14}$$

where DS is the label obtained by the backdoor model detecting the poisoned samples. The total number of successful attacks is recorded as $\sum N_{sa}$. Therefore, the attack success rate is defined as $ASR = \sum N_{sa} / \sum N_{ta} \times 100\%$. Since mAP of the backdoor model and the clean model on clean samples is similar, we lock the object that can be recognized by the model, which can greatly reduce the interference of other factors in the model.

4.1. Experiments of Digital World Attacks

Datasets and Models. The Flir_v2 dataset is released by FLIR Company. We only utilize the thermal infrared images and corresponding annotations, referring to it as $Flir\_v2\_T$, which contain 13460 images and label files. The image size in $Flir\_v2\_T$ is 640×512. The Multi-spectral Object Detection Dataset [27] contains four sub-datasets of RGB, NIR, MIR and FIR, along with ground truth labels provided by an autonomous driving research team at the University of Tokyo. We utilize the FIR sub-dataset containing 7521 thermal infrared images and label files, referred as FIR_Det in the sequel. The image size in FIR_Det is 640×480. We use three mainstream object detection models: YOLO v5, YOLO v3 [38], and Faster RCNN [40] as detectors to verify the attack effectiveness.

Baselines. We follow exactly the same data preprocessing and model training strategies with existing works for the clean model training. The mAP of trained clean model serves as the evaluation baseline for BAF. The results are summarized in Table 1.

Dataset		***Flir_v2_T***		***FIR_Det***	
Class		person	car	person	car
Model (mAP/%)	YOLO v5	79.10	82.50	90.30	93.50
	YOLO v3	84.40	85.80	89.50	92.10
	Faster RCNN	51.65	71.56	79.59	81.54

Table 1. The mAP of person and car in clean samples measured by three models trained on clean datasets $Flir\_v2\_T$ and FIR_Det.

4.1.1 Attack Parameters

We use YOLO v5 detector and $Flir\_v2\_T$ to verify the impact of different attack parameters. The attack setting is that the detector recognizes car as person.

In OAA, the parameters we focus on are **Pixel Value** (p) and **Poisoning Ratio** (q). Unless otherwise specified, the *default parameters* take the following combination: $p = 192$, $q = 20\%$. We list the experimental results in Table 2. The closer the p is to the median, the smaller the difference between the trigger and the object, resulting in a lower BAF of the backdoor model. When the q is increased from 1% to 2%, the attack performance is greatly improved. Therefore, the poisoning ratio should be set above 2%.

Method	Parameter		BAF (%)		ASR (%)
			person	car	
OAA	*Default*		−5.60	−3.40	97.87
	p	255	−1.90	−1.70	97.43
		160	−7.10	−3.40	97.65
		128	−10.40	−5.30	97.78
		64	**−2.20**	**−1.40**	**98.21**
		0	−0.10	−0.90	97.09
	q	15%	−6.10	−3.30	97.63
		10%	−4.80	−2.60	97.30
		5%	**−0.80**	**−0.80**	**92.50**
		2%	−1.00	−1.10	85.33
		1%	0.10	−0.50	52.83
RAA	ar	300	−31.80	−16.40	98.19
		250	−19.50	−8.40	96.50
		200	−6.90	−3.40	96.38
		150	**−1.10**	**−0.90**	**96.55**
		100	−0.30	−0.50	94.15
		50	−0.30	−0.70	77.45

Table 2. The effect of parameters on OAA and RAA.

In RAA, we are more concerned about the **Attack Range** (ar). The p and q are follow *default parameters*. We randomly select a point (160, 206) in the image as the center point of the trigger to fix the trigger location. The attack area is set to a circle with the center point (160, 206) as the center and ar as the radius. We attack the objects whose center point falls within this area. We list the experimental results in Table 2. The smaller the attack range, the less the number of object that can be poisoned (which is why we do not additionally test the poisoning ratio), so ASR will be lower and BAF will be higher. When the attack radius reaches 250, the detection of clean samples will be greatly affected and the attack effect will be reduced.

The additional parametric experiments such as trigger size and relative location, are provided in Appendix B.

4.1.2 Attack Effectiveness

As shown in Table 3, we experiment with two attack methods on the above three models and two datasets to verify attack effectiveness. The attack effects are all chosen as the detector to recognize car as person. In OAA, we set the parameters as $p = 192$ for all datasets. For $Flir\_v2\_T$, $q = 20\%$, while for FIR_Det, $q = 10\%$. In RAA, for $Flir\_v2\_T$ and FIR_Det, the parameter settings are $p = 192$, $q = 20\%$, ar = 150. These parameters are the same for different models. We discover that the closer objects are to the range boundary, the weaker the attack effect becomes. As a result, we set the attack range during inference to be smaller than the range set during training. The results in Table 3 are tested with the attack range of 120. Since Faster RCNN is based on candidate BBox, multi-scale candidate BBoxes can impact the feature extraction of triggers,

Dataset →	*Flir_v2_T*						*FIR_Det*					
Attack Method →	**OAA**			**RAA**			**OAA**			**RAA**		
Model ↓	BAF (%)		ASR (%)	BAF (%)		ASR (%)	BAF (%)		ASR (%)	BAF (%)		ASR (%)
	person	car		person	car		person	car		person	car	
YOLO v5	**−2.90**	**−1.70**	**97.46**	**−0.90**	**−0.90**	**97.44**	**+0.20**	**−0.40**	**97.32**	−0.20	−1.40	96.69
YOLO v3	−1.60	−1.90	96.36	−0.80	−1.20	97.45	−0.50	+0.30	96.65	**−0.90**	**+0.30**	**98.04**
Faster RCNN	−0.41	−0.14	90.01	−0.31	−0.36	84.30	−0.61	−0.70	92.21	−0.84	+1.06	81.61

Table 3. Evaluation results of OAA and RAA on three models and two datasets.

resulting in a weakened attack effect on this model.

Unless otherwise specified, we use the YOLO v5 detector and $Flir\_v2\_T$ to conduct all subsequent experiments. We also verify the attack effectiveness when the attack purpose is car disappearance. In OAA, when the parameter setting is *default parameters*, the BAFs of person and car are -0.5% and -5%, respectively, and the ASR is 98.54%. In RAA, when the parameter setting is *default parameters* and $ar = 150$. The BAFs of persons and cars are $+0.6\%$ and -0.6%, respectively, while ASR is 95.66%.

4.1.3 Temperature Modulated Triggering

For OAA, attack experiments are performed within four different temperature ranges corresponding to different pixel ranges. After testing, we discover that if only triggers within the set temperature range are implanted in the dataset, triggers outside the range still had a high ASR (over 50%). Therefore, we implanted triggers outside the set temperature range for a portion of the data without changing the object label, which is called adversarial triggers. Specifically, 15% of the data is implanted with normal triggers, while 5% of the data is implanted with adversarial triggers. For RAA, we control the attack range using trigger temperatures. Triggers with different pixel values implement RAA with different radii: 80 for pixel value 0 (corresponding to the lowest temperature), 120 for pixel value 128 (corresponding to the average temperature), and 160 for pixel value 255 (corresponding to the highest temperature). We implant triggers with pixel values of 0, 128, and 255, and modify the object labels within the attack radius for 10%, 6%, and 4% of the training data, respectively. The experimental results are presented in Table 4.

More details, along with the Attack Transferability and Comparative Experiments, are provided in Appendix B.

4.2. Experiments of Physical World Attacks

Datasets and Models. We utilize HTI-301 infrared camera (FPA 384×288, NETD < 60mK) for physical experiments, which is the same equipment used in [62]. The size of thermal infrared images produced by this camera is 1420×1080. The object detector is YOLO v5. We use the electric heating device in Figure 5 as the physical trigger. Figure 7 shows illustration images of the deployed trigger in the visible and thermal infrared domains. The device is common enough in real scene to avoid suspicious, and is extremely low in cost.

For OAA, we choose the parking lot as the physical experiment scene. The thermal infrared videos are captured with environment temperature at 32 degrees Celsius. We fix the infrared camera on a moving vehicle. On the same driving route, we record videos with and without triggers, where we randomly selected some cars to place the triggers next to them. After separating the video into frames, we obtain 788 clean images and 472 dirty images. We manually annotate each image with three classes (person, bike, car).

Visible Light **Unheated Trigger** **Heated Trigger**

Figure 7. The trigger for real world deployment.

For RAA, we choose the parking lot and traffic intersection as the physical experiment scenarios. The thermal infrared videos are captured with environment temperature at 30 degrees Celsius. We fix the infrared camera on the side of the road and record videos with and without the trigger that is placed at a fixed location and viewing angle. After separating the video into frames, we obtain 800 clean images and 488 dirty images. We manually annotate each image with four classes (person, bike, car, truck).

Baselines. For the clean images obtained above, we divide the training set and validation set by 9:1. The mAP of the trained clean model is used as the evaluation baseline for BAF. For OAA, the mAP of person (car) in clean samples measured by the clean model is 88.30% (91.80%). For RAA, the mAP of person (car) in clean samples measured by the clean model is 96.70% (98.80%).

Temperature Modulated Triggering. The attack evaluation is shown in Table 4. In the physical world, our methods can effectively attack TIOD. For OAA, the ratio of normal and adversarial trigger implants are 15% and 5%, respectively. For RAA, high-temperature triggers and low-temperature triggers are implanted at ratios of 8% and 12%, respectively. The visualization result is shown in Figure 8.

5. Evaluation of Potential Countermeasures

We empirically evaluate the proposed backdoor attacks when three potential countermeasures are deployed: 1) Pruning[12]; 2) Fine-Pruning[33]; 3) Neural Cleanse[47]. Refer to Appendix C for the more detailed results.

Pruning and Fine-Pruning. These defense methods re-

	Method	Attack Range	Test Range	*Object Misclassification*			*Object Disappearance*		
				BAF (%)		ASR (%)	BAF (%)		ASR (%)
				person	car		person	car	
Digital Attacks	OAA (p)	[0,63]	[0,63] [64,255]	**−1.70**	**−1.80**	**96.75** 5.40	**−1.60**	**−2.30**	**97.19** 5.14
		[64,127]	[64,127] [0,63]∪[128,255]	−3.50	−2.80	95.93 12.83	−0.50	−1.90	90.34 7.57
		[128,191]	[128,191] [0,127] ∪ [192,255]	−8.20	−4.60	96.58 20.05	0.00	−2.30	94.62 21.88
		[192,255]	[192,255] [0,191]	**−2.00**	**−1.50**	**96.61** 6.38	**+0.20**	**−1.10**	**95.41** 5.90
	RAA ($p \setminus ar$)	0 $\setminus\leq$ 80	0 $\setminus\leq$ 80 0 $\setminus>$80	−0.20	−0.40	91.19 4.02	+0.60	−0.40	**95.36** 4.63
		128 $\setminus\leq$ 120	128 $\setminus\leq$ 120 128 $\setminus>$120			89.93 4.39			89.75 5.08
		255 $\setminus\leq$ 160	255 $\setminus\leq$ 160 255 $\setminus>$160			**93.81** 8.55			93.71 7.58
Physical Attacks	OAA (T)	$\leq 26.6°C$	$\leq 26.6°C$ $\geq 36.8°C$	**+8.10**	**+6.20**	**97.83** 5.80	**+9.20**	**+6.00**	**98.38** 8.69
		$\geq 36.8°C$	$\geq 36.8°C$ $\leq 26.6°C$	+3.00	+4.50	97.30 6.52	+4.40	+4.30	95.65 9.42
	RAA ($T \setminus ar$)	$26.6°C \setminus\leq 400$	$26.6°C \setminus\leq 400$ $26.6°C \setminus>400$	−0.30	+0.40	94.32 7.04	+0.20	+0.60	95.60 6.57
		$36.8°C \setminus\leq 600$	$36.8°C \setminus\leq 600$ $36.8°C \setminus>600$			**97.02** 5.13			**97.85** 6.41

Table 4. Experimental results of temperature modulated triggering. The T represents the average temperature of the trigger. The "Attack Range" is the parameter range set during the poisoning phase, and the "Test Range" is the parameter range set during the inference phase.

Figure 8. Examples of OAA and RAA for *car misclassification* in the physical world.

move the backdoor implantation by proving certain neurons. Concretely, we apply them to prune the neurons in the deeper layers of the network, where we vary the number of layers to be pruned from four to eight and the proportion of neurons to be pruned from 20% to 95%. The result shows that while both defense methods can mitigate backdoor attacks, they do so at the cost of decreased accuracy for benign objects. For instance, while capable to lower the ASR to 63.33%, the recognition accuracy for benign person and car also drops to 26.2% (originally 78.3%) and 42.7% (originally 81.7%), respectively.

Neural Cleanse. Neural Cleanse (NC) is a popular defense method against backdoor attacks, which attempts to obtain triggers of each category through reverse engineering. Many subsequent defense methods are based on its ideas. To adapt the NC defense from its original application in the image classification task to our object detection task, we extract the objects from the images in $Flir\_v2\_T$ and label them separately. We take the original dataset and the processed dataset as inputs to NC. NC uses L_1 norm to compute masks and anomaly index to identify toxic objects. The value of anomaly index greater than 2 is considered as a trigger being detected. For OAA, the anomaly index of car is 1.218572. For RAA, the anomaly index of car is 1.767544. Therefore, NC has not detected our attacks.

6. Conclusion

In this paper, we propose two types of backdoor attacks for TIOD: OAA and RAA. Our attacks successfully compromise detectors in both digital and physical worlds, causing them to misidentify cars as persons or fail to detect the presence of cars. We examine various factors that affect the effectiveness of the proposed backdoor attacks. Our research exposes the security vulnerability of these systems and urges for developing effective defenses.

Acknowledgement. This work is supported by the National Key R&D Program of China (No.2022YFB4501300) and the Fundamental Research Funds for the Central Universities (HUST: No.2023JYCXJJ032).

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNS by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 101–105. IEEE, 2019. 2

[2] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5836–5844, 2019. 2

[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 *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 176–194. IEEE, 2021. 2

[4] Shih-Han Chan, Yinpeng Dong, Jun Zhu, Xiaolu Zhang, and Jun Zhou. Baddet: Backdoor attacks on object detection. In *Computer Vision – ECCV 2022 Workshops*, pages 396–412, Cham, 2023. Springer Nature Switzerland. 4, 13

[5] Jinyin Chen, Haibin Zheng, Mengmeng Su, Tianyu Du, Chang-Ting Lin, and Shouling Ji. Invisible poisoning: Highly stealthy targeted poisoning attack. In *Information Security and Cryptology - 15th International Conference, Inscrypt 2019, Nanjing, China, December 6-8, 2019, Revised Selected Papers*, pages 173–198. Springer, 2019. 3

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[7] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2017. 3

[8] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1148–1156. AAAI Press, 2021. 2

[9] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn NIPS Workshop*, 2011. 3

[10] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Appl. Intell.*, 51(3):1244–1261, 2021. 2

[11] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[12] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 7

[13] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[14] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289, 2019. 1

[15] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Mach. Vis. Appl.*, 25(1):245–262, 2014. 1

[16] Jing Gong, Jianhui Zhao, Fan Li, and He Zhang. Vehicle detection in thermal images with an improved yolov3-tiny. In *2020 IEEE international conference on power, intelligent computing and systems (ICPICS)*, pages 253–256. IEEE, 2020. 2

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[18] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017. 1, 3

[19] Zihan Guan, Lichao Sun, Mengnan Du, and Ninghao Liu. Attacking neural networks with neural networks: Towards deep synchronization for backdoor attacks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 608–618, 2023. 3

[20] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023. 3

[21] Yunze He, Baoyuan Deng, Hongjin Wang, Liang Cheng, Ke Zhou, Siyuan Cai, and Francesco Ciampa. Infrared machine vision and infrared thermography with deep learning: A review. *Infrared physics & technology*, 116:103754, 2021. 2

[22] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 3832–3838. ijcai.org, 2022. 2

[23] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund. The effect of a diverse dataset for transfer learning in thermal person detection. *Sensors*, 20 (7):1982, 2020. 2

[24] Terumi Inagaki and Yoshizo Okamoto. Surface temperature measurement near ambient conditions using infrared radiometers with different detection wavelength bands by applying a grey-body approximation: estimation of radiative properties for non-metal surfaces. *NDT & E International*, 29(6):363–369, 1996. 4

[25] Marina Ivašić-Kos, Mate Krišto, and Miran Pobar. Human detection in thermal imaging using yolo. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, pages 20–24, 2019. 2

[26] Chenchen Jiang, Huazhong Ren, Xin Ye, Jinshun Zhu, Hui Zeng, Yang Nan, Min Sun, Xiang Ren, and Hongtao Huo. Object detection from UAV thermal infrared images and videos using YOLO models. *Int. J. Appl. Earth Obs. Geoinformation*, 112:102912, 2022. 2

[27] Takumi Karasawa, Kohei Watanabe, Qishen Ha, Antonio Tejero-de-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, October 23 - 27, 2017*, pages 35–43. ACM, 2017. 6

[28] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. *CoRR*, abs/2004.06660, 2020. 3

[29] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Trans. Dependable Secur. Comput.*, 18(5):2088–2105, 2021. 3

[30] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *CoRR*, abs/2007.08745, 2020. 4

[31] Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world. *CoRR*, abs/2104.02361, 2021. 1

[32] Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, and Shu-Tao Xia. Few-shot backdoor attacks on visual object tracking. *CoRR*, abs/2201.13178, 2022. 3

[33] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*, pages 273–294. Springer, 2018. 7

[34] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. 1

[35] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, pages 182–199. Springer, 2020. 2, 3

[36] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021. 2

[37] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 6

[39] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. 2

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 6

[41] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11957–11965. AAAI Press, 2020. 3

[42] Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*, 2023. 3

[43] Lichao Sun. Natural backdoor attack on text data. *arXiv preprint arXiv:2006.16176*, 2020. 3

[44] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024. 3

[45] Yuhua Sun, Tailai Zhang, Xingjun Ma, Pan Zhou, Jian Lou, Zichuan Xu, Xing Di, Yu Cheng, and Lichao Sun. Backdoor attacks on crowd counting. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5351–5360. ACM, 2022. 1

[46] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8011–8021, 2018. 3

[47] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 707–723. IEEE, 2019. 7

[48] Jintao Wang, Qingzeng Song, Maorui Hou, and Guanghao Jin. Infrared image object detection of vehicle and person

based on improved yolov5. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 175–187. Springer, 2022. 2

[49] Shuo Wang, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. Backdoor attacks against transfer learning with pre-trained deep learning models. *IEEE Trans. Serv. Comput.*, 15(3):1526–1539, 2022. 3

[50] Xingxing Wei, Jie Yu, and Yao Huang. Physically adversarial infrared patches with learnable shapes and locations. *CoRR*, abs/2303.13868, 2023. 1

[51] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6206–6215. Computer Vision Foundation / IEEE, 2021. 3

[52] Mingfu Xue, Can He, Shichang Sun, Jian Wang, and Weiqiang Liu. Robust backdoor attacks against deep neural networks in real physical world. In *20th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2021, Shenyang, China, October 20-22, 2021*, pages 620–626. IEEE, 2021. 1

[53] Mingfu Xue, Can He, Yinghao Wu, Shichang Sun, Yushu Zhang, Jian Wang, and Weiqiang Liu. PTB: robust physical backdoor attacks against deep neural networks in real world. *Comput. Secur.*, 118:102726, 2022. 1

[54] Li Yang, Zhen Yang, et al. *Principle and technology of infrared thermography temperature measurement*. Beijing: Science Press, 2012. 4

[55] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 2041–2055. ACM, 2019. 2

[56] Zenghui Yuan, Yixin Liu, Kai Zhang, Pan Zhou, and Lichao Sun. Backdoor attacks to pre-trained unified foundation models. *arXiv preprint arXiv:2302.09360*, 2023. 3

[57] Zenghui Yuan, Pan Zhou, Kai Zou, and Yu Cheng. You are catching my attention: Are vision transformers bad learners under backdoor attacks? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24605–24615. IEEE, 2023. 2

[58] Zhiyuan Zhang, Lingjuan Lyu, Weiqiang Wang, Lichao Sun, and Xu Sun. How to inject backdoors with better consistency: Logit anchoring on clean data. *arXiv preprint arXiv:2109.01300*, 2021. 3

[59] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14431–14440. Computer Vision Foundation / IEEE, 2020. 1

[60] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12993–13000, 2020. 3

[61] Ce Zhou, Qiben Yan, Yan Shi, and Lichao Sun. {DoubleStar}:{Long-Range} attack towards depth estimation based obstacle avoidance in autonomous systems. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1885–1902, 2022. 3

[62] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3616–3624. AAAI Press, 2021. 2, 4, 7

[63] Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13317–13326, 2022. 1