Tag: THU-AM-300

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 Feng2,3,
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

Paper ID: 10229

Background

- Thermal infrared object detection (TIOD) combines object detection technology with thermal infrared imaging technology, allowing it to recognize objects captured using infrared thermal radiation imaging.

- 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 .

- 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.

- Backdoor attacks pose a serious security threat to deep neural networks (DNNs) due to their stealthiness.

[1] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. Mach. Vis. Appl., 25(1):245–262, 2014.
[2] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. CoRR, abs/1708.06733, 2017.

Motivation

The security vulnerabilities of TIOD remain largely unexplored and current efforts are focused merely on adversarial attacks rather than backdoor attacks.

- Zhu et al. design adversarial patterns and manufacture an adversarial shirt made of aerogel material [3].
- Wei et al. introduce the method of aggregationregularization to optimize the adversarial infrared patch, making the patch easier to implement physically [4].

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.

[3] 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.
[4] Xingxing Wei, Jie Yu, and Yao Huang. Physically adver_x0002_sarial infrared patches with learnable shapes and locations. CoRR, abs/2303.13868, 2023.

Challenge

- Unlike RGB images with three channels, thermal infrared images have only a gray-scale channel and contain less texture information. The channel information available for backdoor attacks on TIOD is significantly less than that for backdoor attacks on VLOD.

- 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.

Question: Can we design effective backdoor attacks on TIOD by utilizing their unique properties compared to VLOD?

Threat Model

Attacker’s Goal

- The first goal is for stealthiness purpose to ensure that the backdoored TIOD can still properly identify the objects in clean samples.

- 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.

Attacker’s Capability

- 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.

Trigger Design

- Existing backdoor attacks for VLOD rely on triggers designed based on color differences [5]. When applied to the thermal infrared domain, as shown in Figure 1, such triggers will appear as grayscale patterns.

- Ordinary materials cannot fully utilize the temperature-sensitive characteristics of thermal infrared modality. As shown in Figure 2, cotton sheets and plastic sheets can only show one morphology in the thermal infrared domain, which cannot meet our attack requirements.

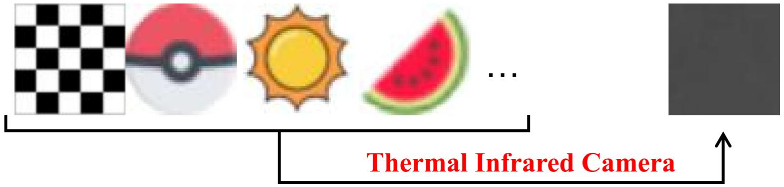


Figure 1. RGB triggers mapped to the thermal infrared domain.

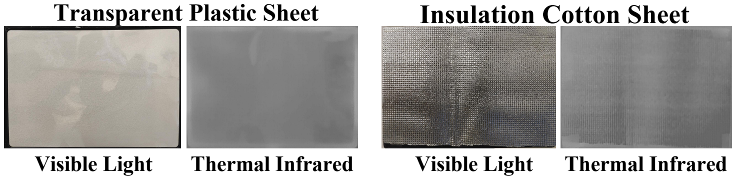


Figure 2. The morphology of ordinary materials in the thermal infrared domain.

[5] 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.

Trigger Design

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

$$p = 1.4221*10^{-4}*T^{4} - 15.4760 \tag{1}$$

- Based on Equation (1), we can simulate digital triggers using pixel blocks. Then, the physical trigger is designed as an electric heating device consisting of an electric heater and a signboard, as shown in Figure 4. In the thermal infrared domain, this trigger can exhibit different morphologies at different temperatures, as shown in Figure 5.

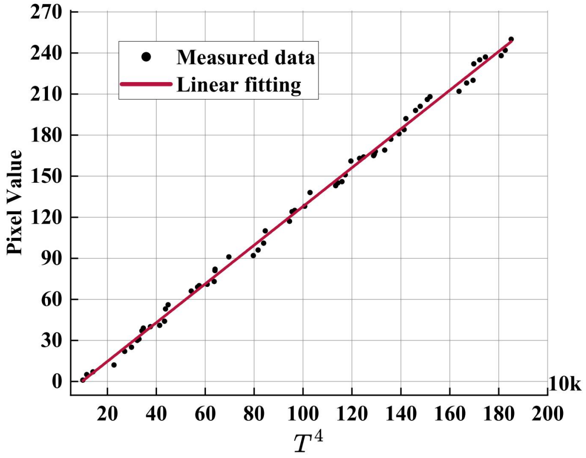


Figure 3. Function fitting of p − T⁴.

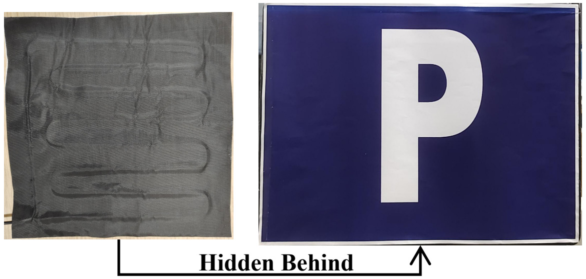


Figure 4. Physical electrothermal trigger design.

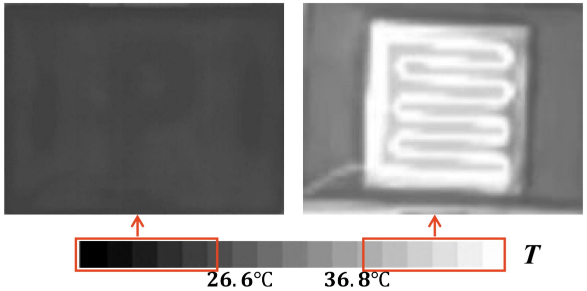


Figure 5. The trigger in the thermal infrared camera.

Overview

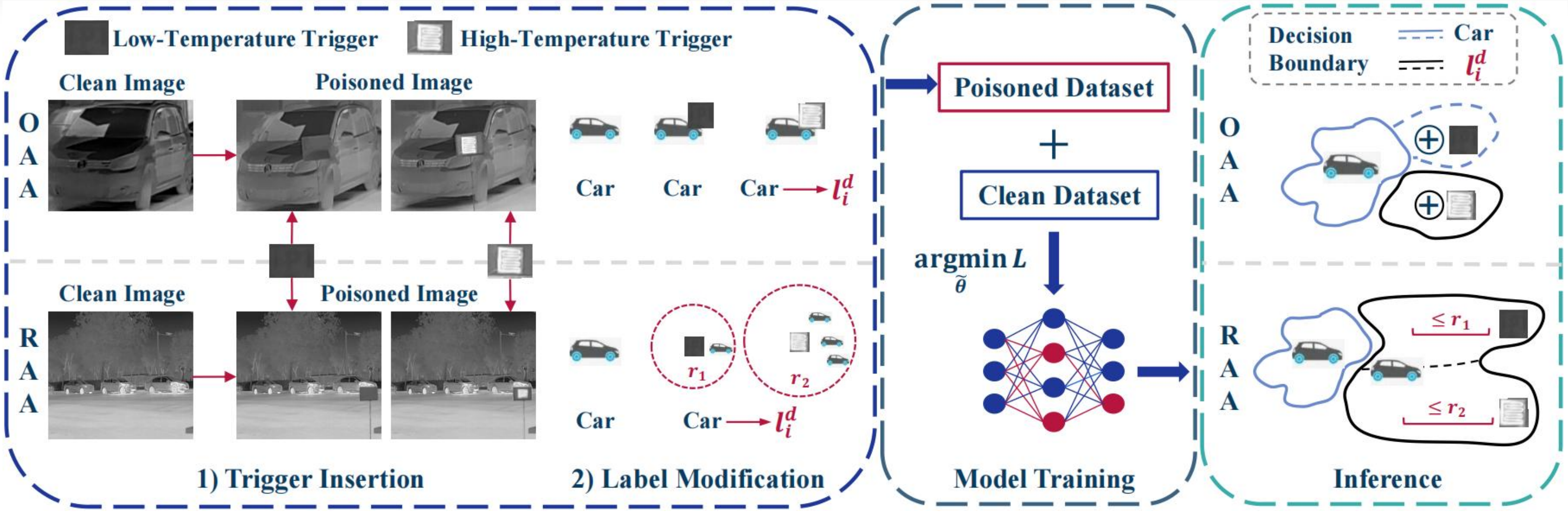


Figure 6. Overview of our proposed attacks.

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.

- Object-Affecting Attack (OAA). OAA implants a trigger in the object and modifies the object's label.
- Range-Affecting Attack (RAA). RAA implants a trigger in an image and modifies the labels of objects of a certain category within a specified range from the trigger.

Label modification method:

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

Experiment Setting

- Evaluation Metrics: Benign Accuracy Fluctuation (BAF) and 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.

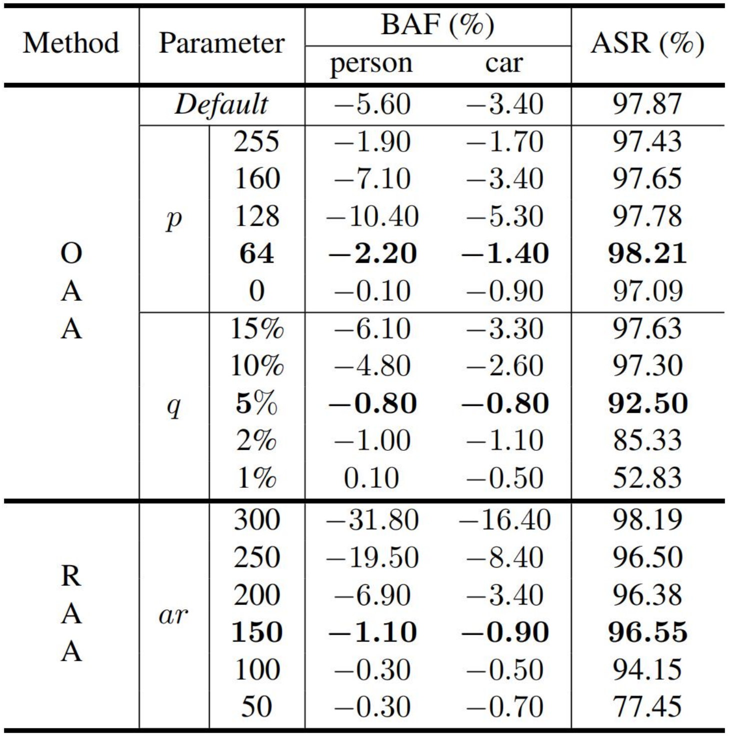
- Datasets: The thermal infrared images and corresponding annotations in the Flir_v2 dataset [6], referred as Flir_v2_T; The FIR sub-dataset in the Multi-spectral Object Detection Dataset [7], referred as FIR_Det.

- Models: YOLO v5[8], YOLO v3[9], and Faster RCNN[10].

[6] https://www.flir.com/oem/adas/adas-dataset-form/
[7] 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.
[8] https://github.com/ultralytics/yolov5
[9] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. CoRR, abs/1804.02767, 2018.
[10] 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.

Digital Experiment Result

Attack Parameters



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

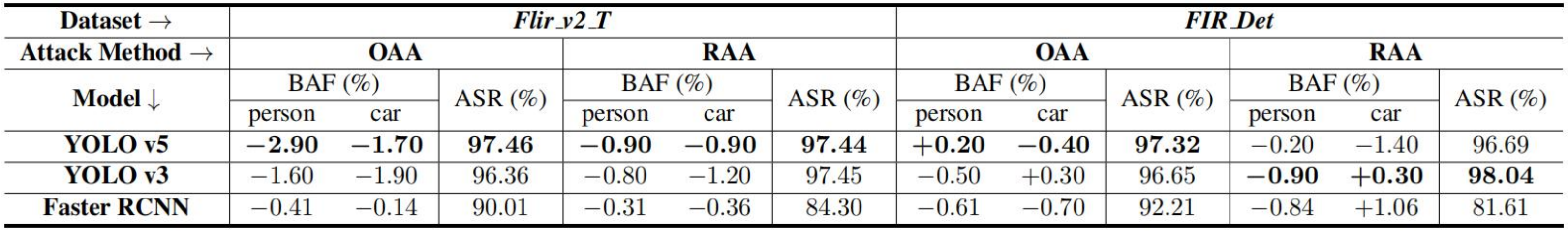
Table 1. The effect of parameters on OAA and RAA.

Default parameters: p=192, q=20%.

- Pixel Value (p). 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.
- Poisoning Ratio (q). When the q is increased from 1% to 2%, the attack performance is greatly improved. Therefore, the poisoning ratio should be set above 2%.
- Attack Range (ar). 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.

Digital Experiment Result

Attack Effectiveness



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

- Our two backdoor attacks are effective on two datasets and three models. Faster RCNN is based on candidate BBox, multi-scale candidate BBoxes can impact the feature extraction of triggers, resulting in a weakened attack effect on this model.

- 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.

Physical Experiment Deployment

Physical Experiment Deployment

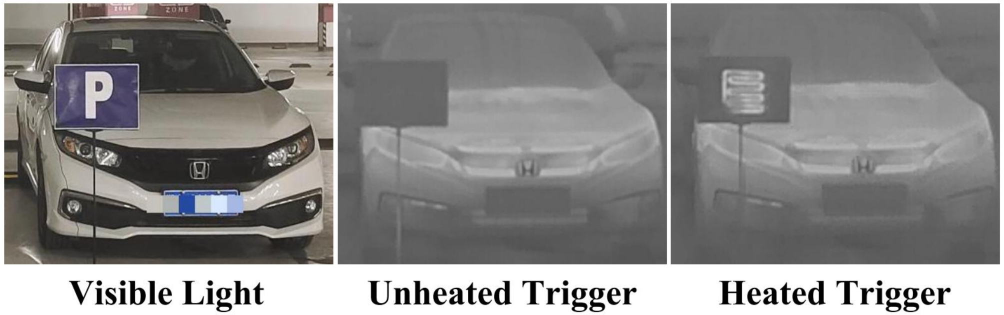


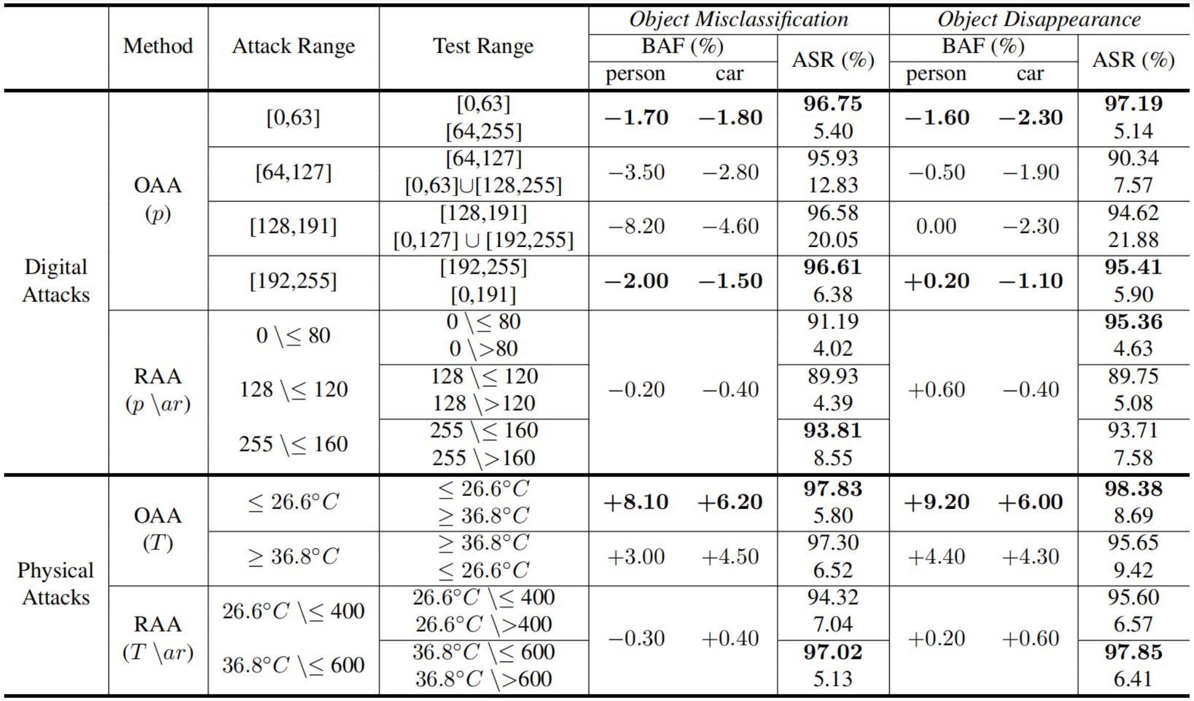
Figure 7. The trigger for real world deployment.

We utilize HTI-301 infrared camera for physical experiments.

- For OAA, we choose the parking lot as the physical experiment scene. 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.

- For RAA, we choose the parking lot and traffic intersection as the physical experiment scenarios. 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.

Temperature Modulated Triggering



	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 \ ar)	0 \≤ 80	0 \≤ 80 0 \>80	−0.20	−0.40	91.19 4.02	+0.60	−0.40	**95.36** 4.63
		128 \≤ 120	128 \≤ 120 128 \>120			89.93 4.39			89.75 5.08
		255 \≤ 160	255 \≤ 160 255 \>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 \ ar)	$26.6°C$ \≤ 400	$26.6°C$ \≤ 400 $26.6°C$ \>400	−0.30	+0.40	94.32 7.04	+0.20	+0.60	95.60 6.57
		$36.8°C$ \≤ 600	$36.8°C$ \≤ 600 $36.8°C$ \>600			**97.02** 5.13			**97.85** 6.41

Table 3. Experimental results of temperature modulated triggering.

- OAA. In the digital environment, attack experiments are performed within four different temperature ranges corresponding to different pixel ranges. In the physical environment, the backdoor attack on the object is achieved through specific temperature ranges.

- RAA. In the digital environment, we control the attack range using pixel values of the trigger. In the physical environment, we control the attack range using temperatures of the trigger.

Temperature Modulated Triggering

Visualization

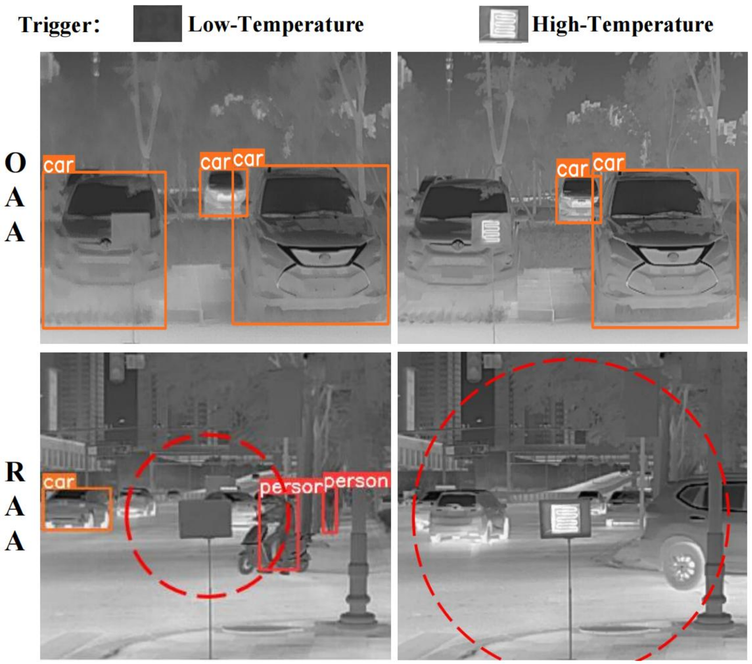


Figure 8. Examples of OAA and RAA for car disappearance in the physical world. The affecting range of RAA is marked as the red circle.

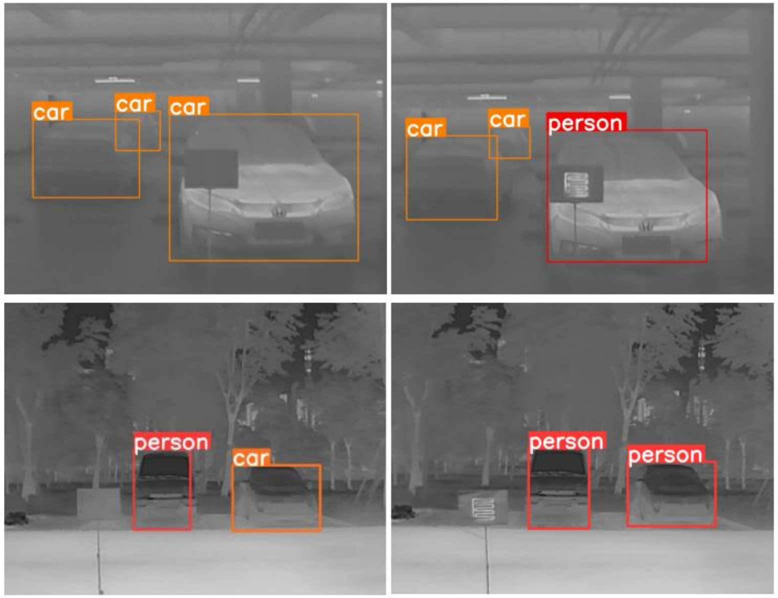


Figure 9. Examples of OAA and RAA for car misclassification in the physical world.

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.

Resistance to Pruning and Fine-Pruning

Pruned Network Layers	Ratio of Pruned Neurons	Pruning			Fine-Pruning		
		BA (%)		ASR (%)	BA (%)		ASR (%)
		person	car		person	car	
21-24	50%	77.30	80.90	95.41	19.30	25.20	9.24
	80%	77.40	80.70	95.49	23.00	33.10	22.11
	95%	77.40	79.90	95.49	23.20	30.10	12.37
19-24	50%	78.10	81.10	95.57	27.00	38.20	19.27
	80%	74.80	75.30	95.47	24.70	36.80	29.42
	95%	74.40	74.50	95.42	22.70	33.90	13.28
17-24	50%	**26.20**	**42.70**	**63.33**	23.70	34.80	65.44
	80%	10.00	14.90	0.00	21.30	34.50	19.65
	95%	1.50	0.10	0.00	24.90	29.40	14.14

Table 4. Evaluation results of Pruning and Fine-Pruning.

- Purning[11]. We prune the back-end network and gradually increase the number of pruned layers. When 50% of the neurons from the 17th to the 24th layers were cut, the ASR dropped to 63.33%. However, the recognition accuracy for benign person and car also drops to 26.2% (originally 78.3%) and 42.7% (originally 81.7%), respectively.

- Fine-Purning[12]. We use the clean dataset to fine-tune the model obtained in Pruning for 20 rounds. When the ASR drops to a relatively low level, the model's recognition accuracy for clean samples also significantly decreases.

[11] 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.

[12] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine_x0002_pruning: Defending against backdooring attacks on deep neural networks. In Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Her_x0002_aklion, Crete, Greece, September 10-12, 2018, Proceedings, pages 273–294. Springer, 2018.

Resistance to Neural Cleanse

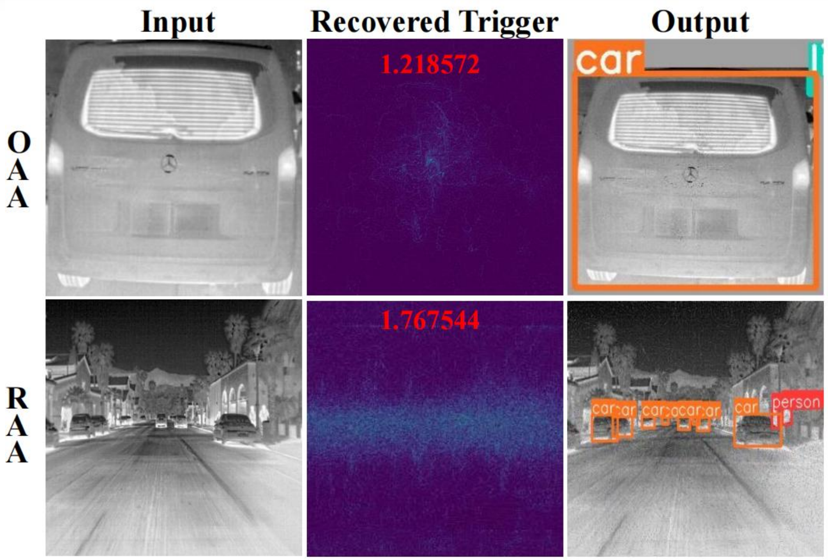


Figure 10. Evaluation results of Neural Cleanse.

- Neural Cleanse (NC) [13]. The Input is clean images fed to NC. The red characters are the anomaly indices (value > 2 considered as trigger detected) detected by NC for the attacked label. The Output is detection results of the backdoor model on images with recovered triggers injected. The anomaly indices obtained for OAA and RAA are both less than 2.

[13] 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.

Conclusion

- 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.

Thank you!