Supplementary Material for Physical Adversarial Clothing Evades Visible-Thermal Detectors via Non-Overlapping RGB-T Pattern

Xiaopei Zhu[1 *] Guanning Zeng[1 *] Zhanhao Hu[2] Jun Zhu[1,3 †] Xiaolin Hu[1,3,4 †]
[1]Department of Computer Science and Technology, BNRist, Tsinghua University
[2]University of California Berkeley
[3]IDG/McGovern Institute for Brain Research, Tsinghua University
[4]Chinese Institute for Brain Research (CIBR)

1. Supplementary Video for RGB-T Physical Attacks

See *Supplementary Video* "Demo Video.mp4" for the demo video of our RGB-T physical attacks.

2. Details for Building 3D RGB-T Models

To construct a 3D RGB-T model, we develop a method for extending a 3D RGB model into an aligned 3D RGB-T model based on a previous thermal 3D modeling approach [22]. Initially, we utilized publicly available 3D RGB human and clothing models [7] as the foundation. Next, we take a clothing model (Fig. S1) as an example to illustrate our method.

The challenge in building an aligned 3D RGB-T model lies in generating an thermal "skin" that aligns with the 3D mesh model (Fig. S1(c)). To address this, we first unfold the faces of the 3D mesh model into a 2D faces map (Fig. S1(d)) and organize it into different regions, such as the back and arms, using Maya software. Next, we capture real thermal images of clothing using an thermal camera and process them to align with the faces map, producing an aligned thermal texture map (Fig. S1(f)). This process ensures that the real thermal texture is properly aligned with the 3D mesh model, and the final rendered 3D RGB-T models are shown in Fig. S1(a) and S1(e).

We observe that cropped thermal images may not perfectly align with the faces map. To address this issue, we utilize Photoshop's distortion function to fine-tune the cropped thermal images for better alignment. In some cases, the captured thermal images may only contain partial regions corresponding to the faces map. When this occurs, we capture images from different angles and stitch them together to create a complete thermal texture map.

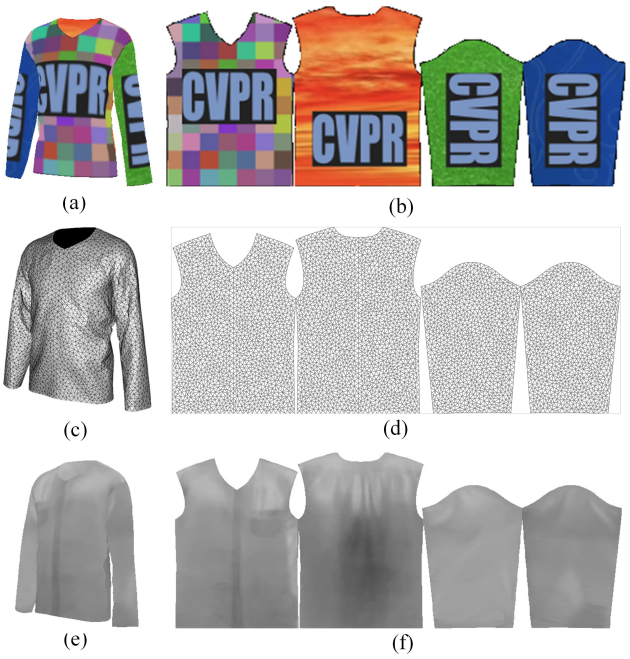

Figure S1. Construction of 3D RGB-T model. (a) 3D RGB model. (b) RGB texture. (c) 3D mesh. (d) Reorganized faces map. (e) 3D thermal model. (f) Thermal texture collected from real world.

Using the above approach, we obtain a fully aligned thermal texture image for the 3D RGB-T model. Finally, we employ PyTorch3D renderer to map both the RGB texture and thermal texture onto the 3D mesh surface, resulting in a 3D RGB-T model.

We further observed that the thermal characteristics of clothing vary with time and location. To simulate these changing thermal properties, we captured thermal images of the same clothing at different times (day and night) and in different locations (indoor and outdoor). Using the method described above, we constructed 3D thermal clothing models for various scenarios, as shown in Fig. S2. During both

*Equal contribution.
†Corresponding authors.

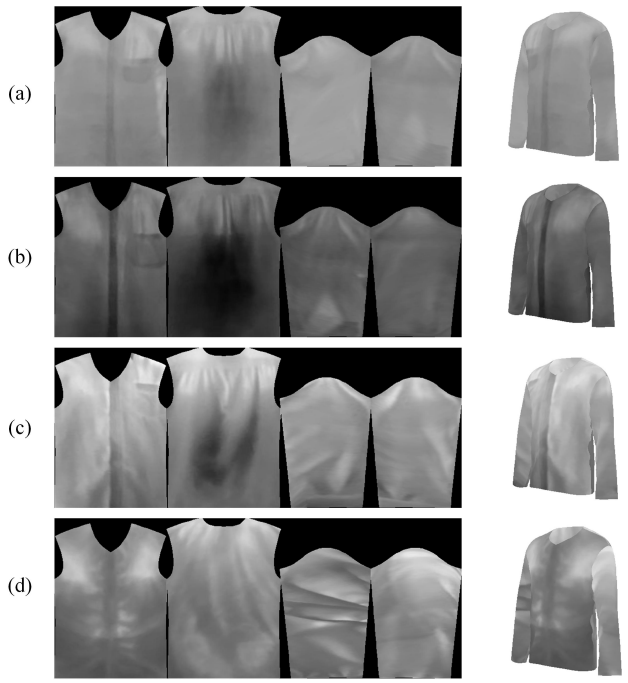

Figure S2. Different thermal characteristics of clothes at (a) day indoors, (b) day outdoors, (c) night indoors, and (d) night outdoors.

the optimization and testing processes, we randomly switch the texture maps of the 3D thermal clothing model to simulate the thermal characteristics of the clothing under different times and locations.

We captured 10 sets of human thermal textures using an thermal imaging camera FLIR T560, covering various environments such as indoor and outdoor settings during both day and night. During the 3D rendering process, visible light is randomly selected from point light sources, directional light sources, and ambient light sources. Infrared light is fixed as an ambient light source to simulate the thermal radiation emitted by the human body in all directions. We increased the brightness of the ambient light source in the thermal rendering to make it close to the thermal imaging effect caused by the high temperature of the human body.

For each sample, we randomly selected the azimuth from 0 to 360 degrees, the elevation from -10 to 20 degrees, and the distance scale from 1.5 to 3.5 for RGB-T joint rendering. After rendering, the human figures were randomly placed within the aligned FLIR background with $x \sim U(-1, 1)$ and $y \sim U(-0.2, 0.2)$, followed by object detection. Considering that some images in the FLIR dataset contain other individuals, their detection results need to be excluded. Therefore, we set an Intersection over Union (IoU) threshold of 0.6 to exclude detection boxes that fall outside the placement range of our rendered human figures, and then took the highest detection score as the loss value.

3. Details for Experimental Settings of Digital Attacks

When conducting spatial discrete-continuous optimization (SDCO), we set the learning rate to 0.01, the batch size to 2, and initialized the attack with $[r_i^{(\mathrm{V})}, g_i^{(\mathrm{V})}, b_i^{(\mathrm{V})}] = U(0, 1)$, $\tilde{p}_i = U(0.5 - 0.01, 0.5 + 0.01)$. The number of training steps was set to 10k steps for the Prob-E, Prob-M, and Prob-L models, and 20k steps for the YOLOv11 model. We set the α of the SRD to 0.7 and the pixel size to 10×10. The texture of shirt was divided into 86×34 pixels, and the texture of pants was divided into 70×48 pixels.

All experiments were performed on RTX3090 GPU with Ubuntu 22.04, CUDA 11.8 and Pytorch 1.13. A single experiment requires around 20GB of GPU memory, takes about 5 hours for every 10k training steps.

After optimization, we rendered the adversarial textures onto the 3D human models (Fig. S4) and placed them in the background images of the test set. We then input the images to RGB-T detectors and evaluate the attack performance both in the white-box and black-box setting. We also compare our method with simple baselines and previous RGB-T attack methods. The results are shown in Tab. 1 and Tab. 5. Fig. S3 shows a set of visual examples.

The results indicates that our method effectively attacked RGB-T detectors with different fusion architectures in the digital world, outperforming simple baselines and previous RGB-T attack methods.

4. Analysis of the Key Parameter of SRD

We further analyzed the impact of the key parameter of SRD, mask probability α, on the ASR. The results are shown in Fig. S5. We observed that highest average ASR was achieved when α was set to 0.7. Therefore, we set $\alpha = 0.7$ in our experiments.

It is worth noting that, α serves as a balancing factor between the optimization parameters of the visible and thermal modalities. When α is too low, the number of trainable parameters in the thermal modality significantly exceeds that in the visible modality, causing the SDCO algorithm to focus more on optimizing the thermal modality than the visible modality. Conversely, when α is too high, SDCO focuses more on optimizing the visible modality. This imbalance in optimization between the two modalities can ultimately degrade the overall attack effectiveness.

5. Effect of Pixel Size

Another hyperparameter that affects the ASR is the pixel size of the adversarial pattern. We tested the ASRs of adversarial clothing patterns with pixel sizes of 5×5, 10×10,

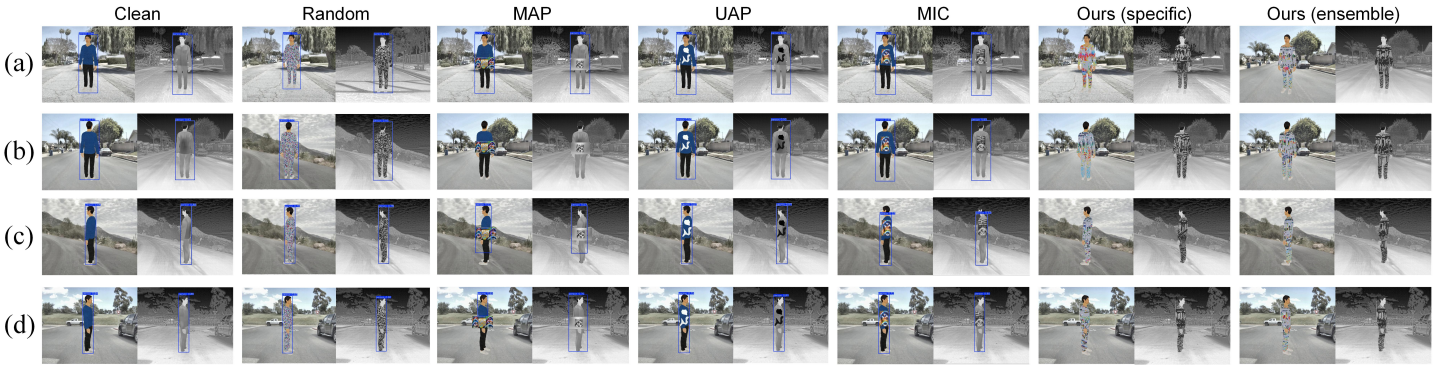

Figure S3. Comparison of different RGB-T attack methods targeting (a) Prob-E, (b) Prob-M, (c) Prob-L, (d) YOLOv11 across diverse scenes in the digital world.

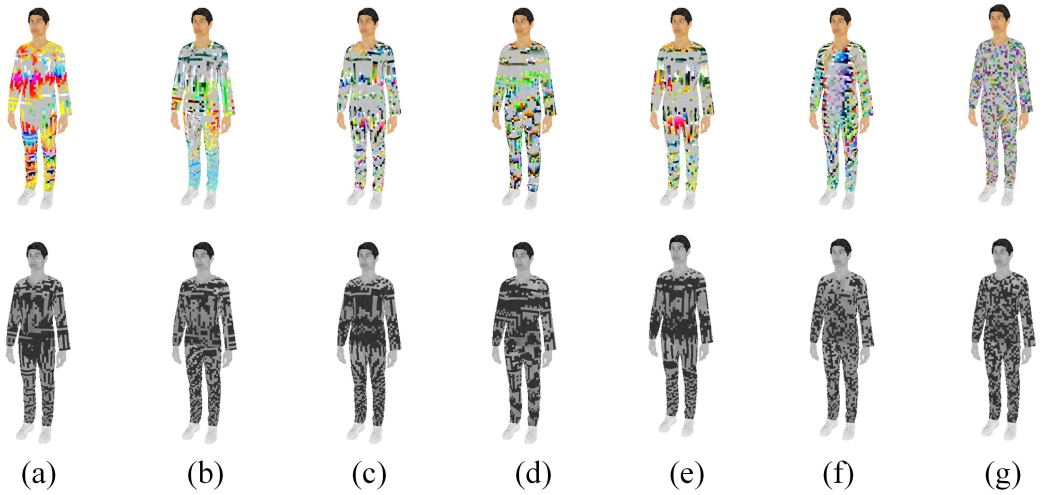

Figure S4. 3D RGB-T models for (a) Prob-E, (b) Prob-M, (c) Prob-L, (d) Ensemble Detector, (e) YOLOv11, (f) Deformable DETR, and (g) Random RGB-T Pattern.

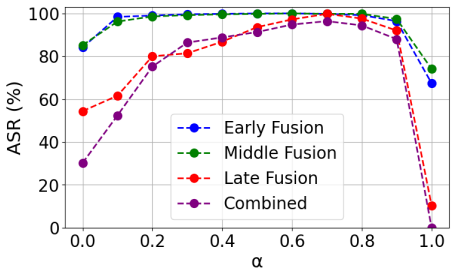

Figure S5. Effect of the parameter α

Table S1. Effect of pixel size

Method	ASR (%) ↑				
	Early	Mid	Late	Combined	
				RGB	T
5×5	99.8	99.6	99.2	86.4	85.6
10×10	100.0	100.0	99.8	98.4	98.2
15×15	100.0	99.8	99.0	94.2	94.6
20×20	99.8	99.8	97.0	87.0	88.4

15×15, and 20×20, targeting RGB-T detectors with different fusion architecture. The results are shown in Tab. S1. We found that adversarial clothing patterns with a pixel size of 10×10 achieved the best performance among these.

6. More Examples for RGB-T Physical Attacks

We provide photos of our adversarial RGB-T clothes in Fig. S7 and additional examples of RGB-T physical attacks in Fig. S6. The captured scenes include both indoor and outdoor environments, spanning different times of the day, including morning, noon, afternoon, and nightfall. Volunteers

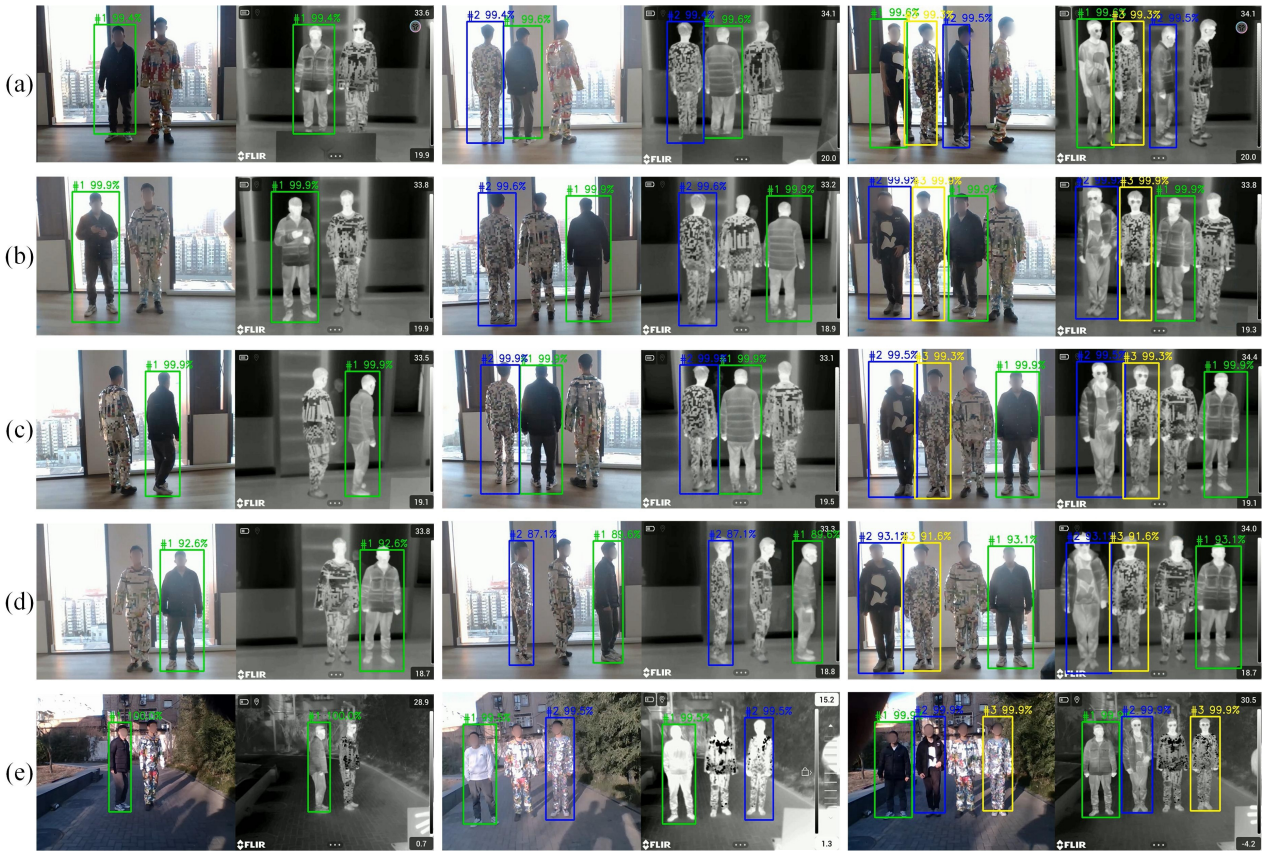

Figure S6. Visualization of physical RGB-T attacks targeting (a) Prob-E, (b) Prob-M, (c) Prob-L, (d) YOLOv11, and (e) Ensemble RGB-T detector at different angles, distances, time, and locations. Facial areas are blurred for privacy reasons.

participated in groups of two, three, or four, wearing our RGB-T adversarial clothes (Fig. S7(a-f)), random-pattern RGB-T clothing(Fig. S7(g)), ordinary clothing, or holding a UAP patch.

The tested RGB-T detectors include both white-box models, such as an early-fusion detector Prob-E [2], a mid-fusion detector Prob-M [2], a late-fusion detector Prob-L [2], and independent RGB-T detectors YOLOv11(RGB) and YOLOv11(T) [9], as well as unseen black-box models, including an unseen early-fusion model RPN-E[12], an unseen mid-fusion model AR-CNN[18], an unseen late-fusion model RPN-L[12], and unseen independent RGB-T detectors D-DETR(RGB) and D-DETR(T)[19]. The camera angles cover a full 0-360° range, and the capture distances range from 2 to 15 meters.

Experimental results indicate that our adversarial clothing successfully evades multiple white-box and black-box RGB-T detectors across various scenes, angles, and distances, consistently outperforming the baseline methods.

7. Attack Transferability in the Physical World

We tested the attack transferability of our RGB-T adversarial clothes to unseen RGB-T detectors in the physical world. The RGB-T detectors involved in the optimization process include an early-fusion detector Prob-E [2], a mid-fusion detector Prob-M [2], a late-fusion detector Prob-L [2], and independent RGB-T detectors YOLOv11(RGB) and YOLOv11(T) [9], and an ensemble model of the aforementioned models. In the testing process, the models, in addition to the ones listed above, also include an unseen early-fusion model RPN-E[12], an unseen mid-fusion model AR-CNN[18], an unseen late-fusion model RPN-L[12], and unseen independent RGB-T detectors D-DETR(RGB) and D-DETR(T)[19]. The results of the physical experiments are shown in Tab. S2. Please note that we calculated the average ASR of both modalities for independent RGB-T detectors in Tab. S2.

The results indicate that our method can transfer to a variety of unseen RGB-T detectors in the physical world. More importantly, our fusion stage ensemble method effectively improves the ASRs against unseen RGB-T detectors compared to patterns optimized for a single model. This

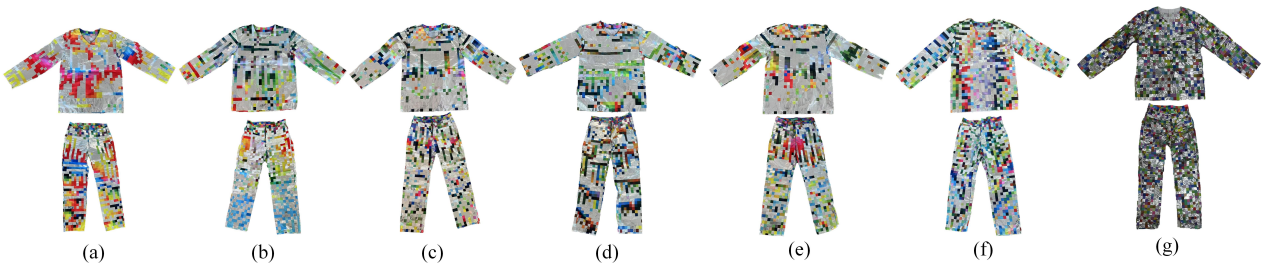

Figure S7. Physical RGB-T clothes for (a) Prob-E, (b) Prob-M, (c) Prob-L, (d) Ensemble Detector, (e) YOLOv11, (f) Deformable DETR, and, (g) Random Pattern.

Table S2. Transferability in the physical world. The numbers are ASRs.

Train \ Test	Prob-E[2]	Prob-M[2]	Prob-L[2]	YOLOv11[9]	RPN-E[12]	AR-CNN[18]	RPN-L[12]	D-DETR[19]
Prob-E[2]	83.6	71.4	21.3	1.9	47.3	51.1	48.1	47.5
Prob-M[2]	57.8	87.8	13.4	3.4	61.4	47.9	58.2	48.1
Prob-L[2]	61.6	77.4	78.4	0.3	75.1	46.2	65.2	41.2
YOLOv11[9]	50.5	55.3	32.3	66.9	39.7	33.9	55.3	27.4
Ensemble	77.6	84.3	78.2	68.8	71.4	55.6	59.0	50.0

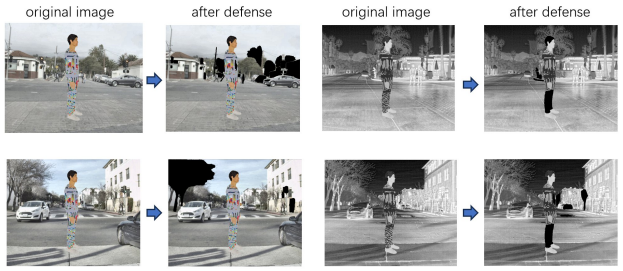

Figure S8. Failure cases of defense. Black segments denote adversarial regions identified by defense algorithms.

suggests that we can just use one single clothing to attack unseen RGB-T detectors with different fusion architectures in the physical world.

8. Details for Defense Methods

We evaluated the effectiveness of eight typical adversarial defense methods against our attack methods. These included five traditional defense techniques: Adversarial Training[3], Total Variance Minimization [1], Bit Squeezing[16], JPEG Compression[4], and Pixel Mask[4], along with three state-of-the-art methods specifically designed for defending against adversarial attacks on object detectors: PAD[8], NAPGuard[15], and Jedi[13].

For Adversarial Training, we began by rendering 200 images with adversarial textures and collecting their corresponding labels. These images were then added to the original FLIR-aligned dataset for fine-tuning RGB-T detectors. We fine-tuned the Prob-E, Prob-M, Prob-L, and YOLOv11 models individually for 10 epochs, with a learning rate of 0.001. The fine-tuned model weights were subsequently used as the target for our attacks. In the case of Pixel Mask defense, we randomly selected an 80×80 region on both the shirt and trousers textures, simultaneously removing the thermal materials and visible light adversarial textures from these areas. For Bit Squeezing, we reduced the bit depth of both the 8-bit visible light and thermal adversarial textures to 7-bit. Regarding Total Variance Minimization and JPEG Compression, we utilized modules from the Adversarial Robustness Toolbox library to compress or blur the adversarial textures. As for PAD, NAPGuard, and Jedi, we directly implemented their open-source codes, which attempt to verify and eliminate our adversarial clothing on simulated humans.

Tab. S3 shows the results. It indicates that although these

Table S3. Evaluation of Defense Methods

Method	ASR drop(%) ↓				
	Early	Mid	Late	Combined	
				RGB	T
AT[3]	23.4	18.4	21.0	25.4	24.6
TVM[1]	17.0	14.6	9.8	14.6	18.0
BIT[16]	9.8	7.2	8.6	12.4	14.2
JPEG[4]	11.2	10.0	15.4	16.8	19.6
PM[4]	21.0	17.4	24.2	12.0	10.2
PAD [8]	17.2	12.4	21.2	25.0	22.0
NAPGuard [15]	15.4	11.0	22.4	28.6	17.2
Jedi [13]	22.4	19.8	29.8	27.6	28.2

methods had some defense effects, the ASRs of our method after defense still achieved at least 70%, further indicating the effectiveness of our attack approach. Note that our 3D modeling ensures that adversarial clothing can cover larger areas of human body and have more irregular shapes and boundaries compared with 2D adversarial patches. Therefore, even latest methods specialized in defending object detectors cannot precisely locate our adversarial clothing. Fig. S8 illustrates examples of failure cases of defense when applying PAD method.

9. Attack Deformable DETR

To evaluate the attack effectiveness of our method against the transformer-based detector, we attacked a typical detector, Deformable DETR [19], both in the digital and physical world. Following the experimental setup described in Sec. 4.4, we obtained the optimized adversarial clothing, as shown in Fig. S4(f). We then tested its attack performance in the digital world, and the experimental settings were same as Sec. 4.4. The ASR of the adversarial clothing was 99.6% in the digital world. In comparison, the ASR of the random pattern clothing was only 17.8%.

Next, we manufactured adversarial clothes based on the optimized patterns, as shown in Fig. S7(f), and conducted physical experiments following the setup described in Sec. 4.9. The physical experiments indicated that our adversarial clothes successfully evaded Deformable DETR in the physical world, achieving an ASR of 75.4%, while the ASR of random pattern clothes was only 22.0%. These results, together with our previous experiments, indicate that our method is effective against both the CNN-based models (e.g., YOLOv11) and the transformer-based model, highlighting the generality of our approach.

10. Attack E2E-MFD and DAMSDET

Table S4. Evaluation on E2E-MFD and DAMSDET

Method	ASR (%) ↑	
	E2E-MFD	DAMSDET
Clean	0.2	3.6
Random	0.4	9.6
MAP[10]	8.8	17.8
MIC[11]	11.4	14.2
UAP[14]	7.0	12.0
Ours	88.2	94.6

We evaluated our attack method on two recently published RGB-T detectors—an early-fusion detector E2E-MFD [17] and a mid-fusion detector DAMSDET[5]. For a fair comparison, we employed the clean clothing pattern (pure color pattern), random RGB-T pattern (without optimization), and the adversarial patterns generated by previous works including MAP[10], UAP[14], and MIC[11]. The results are shown in Tab. S4. Our method achieved an average ASR of 91.4%, while the ASR for the control group was below 17.8%. This indicates that our approach can effectively attack state-of-the-art RGB-T detectors, outperforming simple baselines and previous RGB-T attack methods.

11. Limitation and Future Work

As the first physical attack that effectively targets RGB–T detectors across all fusion methods, this paper accepts a modest loss in garment naturalness, consistent with early work on single-modality [6, 10, 11, 20, 21], to give research priority on better exposing the vulnerabilities of current RGB–T object detectors. Future work would further focus on how to improve the perceptual naturalness of clothing patterns while remaining effectiveness of attack in multimodal real-world detection.

12. Ethics Statement

Adversarial example techniques should be used carefully. If abused, adversarial attacks may threaten the security of AI systems. However, adversarial attacks also advance AI robustness research by exposing system vulnerabilities and promote the development of more robust and trustworthy AI models.

References

[1] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *Pattern Recognition Letters*, 146:244–251, 2021. 5

[2] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022. 4, 5

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 5

[4] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 5

[5] Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *ECCV 2024*, pages 464–481. Springer, 2024. 6

[6] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13307–13316, 2022. 6

[7] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16975–16984, 2023. 1

[8] Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. Pad: Patch-agnostic defense against adversarial patch attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24472–24481, 2024. 5

[9] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 4, 5

[10] Taeheon Kim, Hong Joo Lee, and Yong Man Ro. Map: Multispectral adversarial patch to attack person detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4853–4857. IEEE, 2022. 6

[11] Taeheon Kim, Youngjoon Yu, and Yong Man Ro. Multispectral invisible coating: laminated visible-thermal physical attack against multispectral object detectors using transparent low-e films. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1151–1159, 2023. 6

[12] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 4, 5

[13] Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4087–4095, 2023. 5

[14] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[15] Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, and Xianglong Liu. Napguard: Towards detecting naturalistic adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24367–24376, 2024. 5

[16] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 5

[17] Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, Daixun Li, Wenbo Huang, Yunsong Li, and Xue Yang. E2e-mfd: Towards end-to-end synchronous multimodal fusion detection. *Advances in Neural Information Processing Systems*, 37:52296–52322, 2024. 6

[18] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4, 5

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4, 5, 6

[20] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3616–3624, 2021. 6

[21] Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in realworld. In *CVPR*, 2022. 6

[22] Xiaopei Zhu, Yuqiu Liu, Zhanhao Hu, Jianmin Li, and Xiaolin Hu. Infrared adversarial car stickers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24284–24293, 2024. 1