ACAttack: Adaptive Cross Attacking RGB-T Tracker via Multi-Modal Response Decoupling

Xinyu Xiang Qinglong Yan Hao Zhang* Jiayi Ma*
Wuhan University, China
xiangxinyu@whu.edu.cn, qinglong_yan@whu.edu.cn, zhpersonalbox@gmail.com, jyma2010@gmail.com

Abstract

The research on adversarial attacks against trackers primarily concentrates on the RGB modality, whereas the methodology for attacking RGB-T multi-modal trackers has seldom been explored so far. This work represents an innovative attempt to develop an adaptive cross attack framework via multi-modal response decoupling, generating multi-modal adversarial patches to evade RGB-T trackers. Specifically, a modal-aware adaptive attack strategy is introduced to weaken the modality with high common information contribution alternately and iteratively, achieving the modal decoupling attack. In order to perturb the judgment of the modal balance mechanism in the tracker, we design a modal disturbance loss to increase the distance of the response map of the single-modal adversarial samples in the tracker. Besides, we also propose a novel spatio-temporal joint attack loss to progressively deteriorate the tracker's perception of the target. Moreover, the design of the shared adversarial shape enables the generated multi-modal adversarial patches to be readily deployed in real-world scenarios, effectively reducing the interference of the patch posting process on the shape attack of the infrared adversarial layer. Extensive digital and physical domain experiments demonstrate the effectiveness of our multi-modal adversarial patch attack. Our code is available at https://github.com/Xinyu-Xiang/ACAttack.

1. Introduction

The adversarial attack on visual object tracking (VOT) [1, 26] aims to mislead the prediction results of the tracker through the generated adversarial disturbance, find the model vulnerabilities, and then promote the security of the tracking model in real-life. Single-modal tracking attack methods have been extensively studied, but with the wide application of multi-modal devices [19, 20, 24], multi-modal trackers are widely used in safety-critical real-world fields such as autonomous driving and urban security [16].

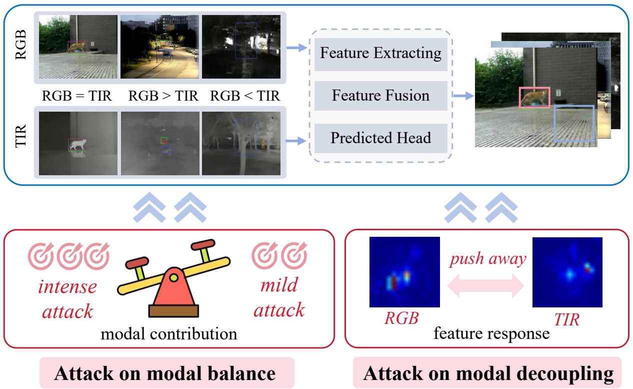


Figure 1. Our attack strategy against RGB-T trackers. An adaptive attack strategy, sensitive to modality, is introduced to alternately and iteratively suppress the modality with a high contribution of shared information. Additionally, a modal disturbance loss is crafted to enlarge the response map distance for single-modal adversarial samples within the tracker.

To address the urgent need to explore the security of trackers, adversarial attack techniques for trackers have emerged in rapid succession, including traditional gradient-based attack approaches and deep-network-based attacks. The former methods [10, 11] utilize hand-crafted parameters and apply many times of gradient ascent to maximize an adversarial loss function for misguiding deep networks. Although it can achieve certain attack effects for specific types of trackers, it is challenging to comprehensively explore and attack the potential vulnerabilities of the different trackers due to the limitations of inflexible pattern design. Nonetheless, the latter one [6] applies tremendous data to train an adversarial patches-generator including flexible architectures and optimization strategies to better automatically search for model weaknesses and realize tracker attacks. Therefore, in comparison with a traditional gradient-based attack approach, deep-network-based paradigm can more automatically and flexibly excavate the security issues within the tracker.

*Corresponding authors.

Although prior efforts for adversarial attack methods are effective in interference trackers, several challenges still need to be addressed. Notably, existing adversarial attack methods [5, 15, 25] on tracking are designed for RGB modality, whereas the methodologies for attacking RGB-T multi-modal trackers are less explored so far. Considering the widespread deployment of multi-modal tracking technology [17, 27, 28] in several safety-critical areas, it is urgent to explore and implement adversarial attacks of multi-modal tracking to understand the potential vulnerabilities of the trackers. However, as shown in Fig. 1, the unique modal coupling and structural design of RGB-T trackers make it a great challenge to successfully find model vulnerabilities. Firstly, due to the modal equilibrium mechanism and coupled multi-modal information, it is difficult to successfully jam the RGB-T tracking model itself. Specifically, the modal balancing strategy in the multi-modal tracker can effectively prevent the attack of adversarial perturbation in a single modal. Secondly, the coupling of multi-modal information can effectively weaken the attacks against the consensus region of the target. Thirdly, the deployment of patches in the physical world is also challenging because the stacked placement of multi-modal patches has a probability of compromising the expression of infrared adversarial shapes, reducing their synergy performance.

Considering these challenges, we propose ACAttack, an adaptive cross attack framework via multi-modal response decoupling. It aims to generate multi-modal adversarial patches to evade RGB-T trackers in both digital and physical domains. Specifically, this framework can gradually and adaptively optimize, discover plenty of rough adversarial samples, and then map them to the high-dimensional adversarial space of different modalities according to the modal response contribution factor, forming multi-modal adversarial patches. Secondly, a modal-aware adaptive attack strategy is introduced to weaken tracker's deep semantic attention to the modality with high common information contribution according to the contribution degree of modal response alternately and iteratively, achieving the modal decoupling attack. When the contributions of two modalities are similar, we design a modal disturbance loss to search the modal imbalance vulnerabilities of the tracker, expand the distance of the response map of the single-modal adversarial samples in the tracker, and perturb the judgment of the balance modal in the tracker. We also design a spatio-temporal joint attack loss to build progressively enlarged pseudo-GT between consecutive frames, which progressively deteriorates the tracker's perception of the target. Thirdly, the design of the shared adversarial shape is deployed to eliminate the interference of visible patches on the expression of infrared adversarial shapes. After the shape is shared, it can not only reduce the consumption of the adversarial shape's inter-modal attack ability but also realize attacks other than texture in the visible modal.

In summary, we make the following contributions:

- We make an innovative attempt to propose an adaptive cross attack framework via multi-modal response decoupling. It can generate multi-modal adversarial patches to mislead RGB-T trackers effectively.
- We develop a novel modal attack flow, in which modal-aware adaptive attack strategy and modal attack constraints alternately disturb the modes with high contribution to achieve modal decoupling and destroy modal balance mechanism of tracker, respectively.
- We design the shape-shared stack strategy to linkage the visible and infrared adversarial shapes, reducing the attack consumption of multi-modal patches mutual deployment in physical scenarios.
- Experimental results show that multi-modal patches can efficiently fool RGB-T trackers in standard RGB-T tracking datasets and real scenes.

2. Related Work

2.1. Visual Object Tracking

Given the tracked object in the first frame, object tracking aims to recognize and locate the object in subsequent frames. Many RGB tracking methods [2, 7, 14, 18, 21] have been proposed and achieved commendable tracking performance. However, RGB sensors struggle to capture objects effectively under challenging conditions such as occlusion and low light, limiting the performance of RGB trackers. To address this, the RGB-T tracking paradigm is introduced, which is not restricted to a single RGB modality but instead integrates the complementary information from both RGB and thermal modalities. This fusion enables more robust tracking capabilities. ViPT [29] introduces a vision prompt tracking framework that leverages the foundational model with strong representation capabilities, enabling interaction between the thermal and RGB modalities through a modality-complementing prompter. BAT [3] proposes a universal bidirectional adapter, which enables mutual prompt between the thermal and RGB modalities and further improves tracking performance. SDSTrack [9] designs a complementary masked patch distillation strategy based on self-distillation learning, which enhances the tracking robustness in extreme weather.

2.2. Adversarial Attacks

Currently, adversarial attacks in the tracking task primarily target RGB trackers. For instance, APYVOT [4] proposes an optimization objective function with a dual-attention mechanism to generate perturbations, disrupting tracking by interfering solely with the initial frame. MTD [8] introduces a maximum textural discrepancy loss function that misleads the visual trackers by decorrelating the template

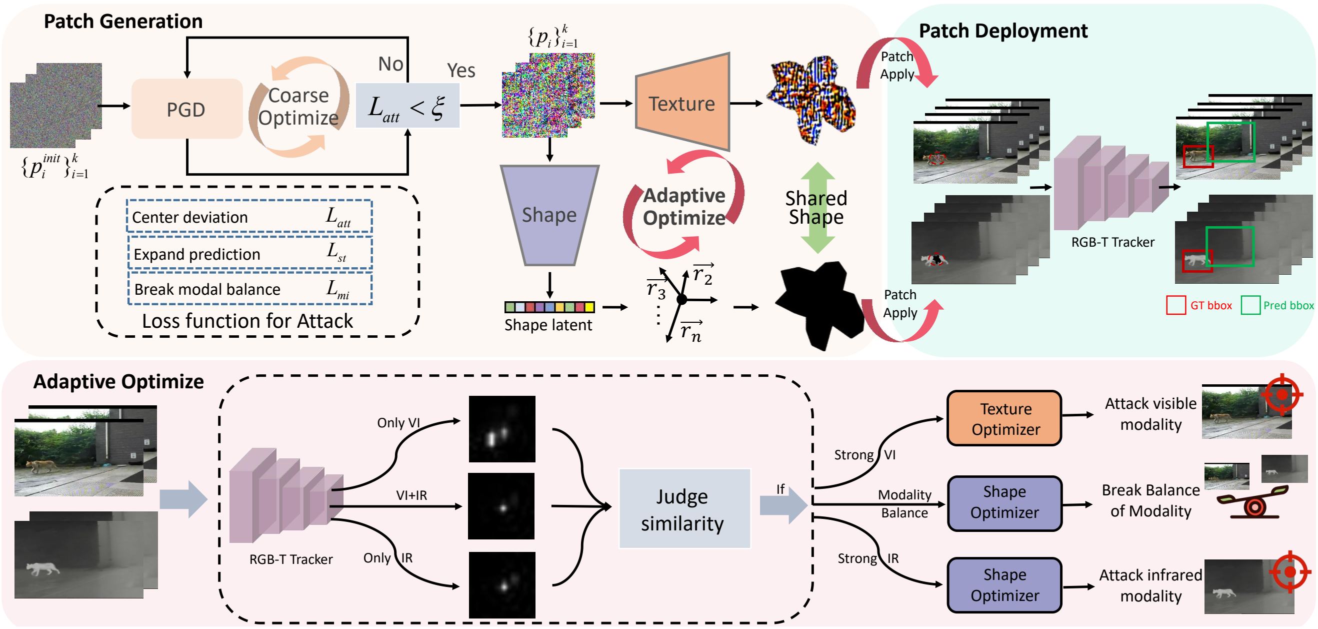


Figure 2. The overall framework of our ACAttack.

and search frame at hierarchical feature scales. These methods, however, fail to disrupt the significant feature enhancement resulting from the interactions between RGB and thermal modalities, which limits their effectiveness against RGB-T trackers. Therefore, it is essential to develop the attack strategy specifically designed for RGB-T tracking.

3. Methodology

3.1. Coarse-to-Fine Modality Attack Framework

With the help of the progressive modality information integration strategy, the multi-modal trackers gradually strengthen the common scene representation and target response, thereby achieving a robust tracking performance superior to that of the single-modal trackers. Consequently, a coarse-to-fine architecture is designed to progressively degrade the modality integration capability of RGB-T models named ACAttack, which can be divided into two stages. The overall architecture of our ACAttack is illustrated in Fig. 2 and Algorithm 1. First, we employ projected gradient descent (PGD) in stage1 to identify a set of adversarial examples $\{p_i\}_{i=1}^{k}$ with sufficient aggressiveness, narrowing the search space for refined attacks and increasing the likelihood of discovering strong adversarial examples, formulated as:

$$\{p_i\}_{i=1}^{k} = PGD(p_i^{init}), \tag{1}$$

where p_i^{init} is randomly initialized with noise patches. The generated patches are subsequently loaded onto the visible image I_{vi} to form a visible adversarial sample I_{vi}^{adv}.This process can be formulated as follows:

$$I_{vi}^{adv} = p_i \odot M + I_{vi} \odot (1 - M), \tag{2}$$

where M is the binary mask for applying adversarial patch. $\odot$ represents the element-wise Hadmard product. The adversarial visible image I_{vi}^{adv} concat with clean infrared image I_{ir} are sent to RGB-T tracker $T(\cdot)$ to predict final bounding box $Bbox_{pred}$ of target, which is expressed as:

$$Bbox_{pred} = T(I_{vi}^{adv}, I_{ir}). \tag{3}$$

We optimize this process by minimizing the conventional attack loss L_{att} relative to the center point, which is defined as:

$$L_{att} = -\|(Cp(Bbox_{pred}) - Cp(Bbox_{gt}))\|_2^2, \tag{4}$$

where Cp denotes the operator to obtain the center point of bounding box $Bbox$. Specifically, when the attack loss L_{att} reaches a predefined threshold, the iteration is halted, and a rough adversarial sample is generated. This process is repeated k times to generate a set of k rough adversarial samples. Considering the different imaging principles of infrared and visible modalities, the multi-modal patches will be specially designed according to their differences in principles. Specifically, the visible modal mainly employs the attack texture to interfere. For the infrared modal, it is difficult to detect the texture, so the adversarial shape is used to attack. Subsequently, the set of adversarial samples generated in the coarse attack stage (stage1) is fed into the subsequent fine attack process (stage2) for further refinement, resulting in the generation of multi-modal adversarial patches with strong attack performance. Finally, through continuous iterative optimization, the multi-mode patch will share the same attack shape, while the visible patch will also possess adversarial texture to confuse the tracker. The fine-grained attack phase targets the modality of the multi-modal tracker and consists of modal decoupling attacks and modal balance interference, which will be detailed in the subsequent sections.

3.2. Modal Decoupling Attack

The RGB-T tracker implicitly couples the contributions of the two modalities, thereby enhancing tracking accuracy. Given the significant role of modal contribution in the tracker, we propose a modal decoupling attack to adaptively diminish the influence of advantageous modalities. Specifically, we use the coarse adversarial samples from the stage1 as input for the stage2, feeding them simultaneously into the adversarial texture generation network G_{tex}^{Adv} and the adversarial shape generation network G_{shape}^{Adv}. For attacking the infrared modality, the rough adversarial sample set from the first stage is encoded into r dimensions via continuous downsampling and an MLP, controlling the adversarial shape. The infrared patch p_{ir} generation process can be expressed as follows:

$$p_{ir} = G_{shape}^{Adv}(\{p_i\}_{i=1}^{k}). \tag{5}$$

The adversarial texture generation network generates adversarial textures to attack the visible modality using residual connections and upsampling [23], which is defined as:

$$p_{vi} = (1 - p_{ir}) \odot G_{shape}^{Adv}(\{p_i\}_{i=1}^{k}), \tag{6}$$

where the visible patch with adversarial textures is p_{vi}.

Subsequently, the modal contribution of the current network input is calculated as the reciprocal of the difference between the response map obtained from single-modal data and the response map from dual-modal input. A larger reciprocal distance indicates that the response maps from dual-modal and single-modal inputs are more similar, suggesting a greater contribution from the current single modality to the tracker. The modal response contribution can be expressed as follows:

$$c_m = \frac{1}{dis(R(m,m), R(vi,ir))}, \tag{7}$$

where c_m represents the contribution value of $m \in \{vi, ir\}$ modal to the tracker. dis stands for the distance function and is used to measure the Euclidean distance between response maps. $R(\cdot,\cdot)$ shows the response map acquired by the tracker under the current input.

Algorithm 1: The ACAttack Algorithm

```
Input: Random patches $p_i^{init}$, parameters $k$,
       $M_{stage1}$, $M_{stage2}$, $\xi$, $\zeta$
Output: Optimized multi-modal patches $p_{vi}, p_{ir}$
Iteration:
    Initialize a random patch $p_i^{init}$;
    $i = i + 1$;
    Iteration:
        Generate $p_i$ through Eq. (1);
        Use Eq. (2) to generate adversarial sample $I_{vi}^{adv}$;
        Calculate $Bbox_{pred}$ using Eq. (3);
        Optimize $PGD(\cdot)$ with Eq. (4);
    Until: $L_{att} < \xi$ or $iter \geq M_{stage1}$
Until: $i \geq k$
Determine $\{p_i\}_{i=1}^{k}$ after optimazation in stage1;
Iteration:
    $iter = iter + 1$;
    Obtain $p_{ir}, p_{vi}$ via Eqs. (5) and (6);
    Apply multi-modal patches $p_{ir}, p_{vi}$ on $I_{ir}, I_{vi}$ ;
    Calculate $c_{vi}, c_{ir}$ using Eq. (7) ;
    Send to Tracker $T(\cdot)$ to predict bounding box;
        if $|c_{vi} - c_{ir}| < \zeta$;
            Optimize $G_{shape}^{Adv}$ with Eq. (9);
        elif $c_{vi} - c_{ir} > \zeta$;
            Optimize $G_{tex}^{Adv}$ with Eqs. (4) and (10);
        elif $c_{ir} - c_{vi} > \zeta$;
            Optimize $G_{shape}^{Adv}$ with Eqs. (4) and (10);
Until: $iter \geq M_{stage2}$
```

To normalize the modal contribution, a softmax operation is applied to the reciprocal distance, yielding the final modal contribution score, formulated as:

$$c_{norm} = softmax(c_{vi}, c_{ir}). \tag{8}$$

Finally, an automatic discriminant attack is executed based on the modal contribution score. As illustrated, when the visible contribution is higher in the input data, only the visible modal is attacked, specifically by optimizing the generation of adversarial textures. When the infrared contribution is higher in the input data, only the adversarial shape is modified to attack the infrared modal, thereby reducing its contribution to the tracker. Given that the tracker employs a modal balance mechanism, the contributions of the two modalities may be similar in certain scenarios, as detailed in the subsequent section.

3.3. Modal Balance Interference

Previous work on tracking attacks has attempted to design explicit attack losses to detect model vulnerabilities, but this approach often fails to account for the inherent characteristics of the model, making it challenging to execute effective attacks. Inspired by the concept of implicit attacks [25] and the multi-modal aggregation properties in

RGB-T tracker [22], we develop a loss function with modal-balanced interference to target multi-modal trackers. In cases where the contributions of infrared and visible modal are similar (i.e., modal balance), the response map of single-modal input closely resembles that of dual-modal input. To disrupt this balance, we extract the response maps of the two single-modal adversarial examples and increase the distance between them. The details are provided as follows:

$$L_{mi} = -\|R(vi_{adv}, vi_{adv}) - R(ir_{adv}, ir_{adv})\|_2^2. \tag{9}$$

Notably, the infrared and visible patches in our method share the same adversarial shape to achieve simultaneous attacks on both modalities. Therefore, under conditions of modal balance, only the adversarial shape is optimized. Additionally, a spatio-temporal joint attack loss L_{st} is employed in conjunction with the modal jamming loss L_{mi} to disrupt the tracker's semantic perception. The specific design is presented in the following formula:

$$L_{st} = \left\| \sum_{i=1}^{s} Bbox_{pred}(w,h) - r_i * Bbox_{gt}(w,h) \right\|_2^2, \tag{10}$$

where s denotes the consecutive $s = 5$ frames extracted from a video. r_i represents the scaling factor over time to construct the pseudo-GT, which is set as $[1.90, 1.95, 2.00, 2.05, 2.10]$.

3.4. Implementation Process in Real-world

After completing the digital domain optimization, the multi-modal adversarial patches require deployment in the real world. However, during real-world deployment, visible and infrared patches are stacked, leading to inevitable interactions between the two modalities, as illustrated in Fig. 3. Specifically, the coverage of visible patches impacts the adversarial shape expression of infrared patches, while the presence of infrared patches hinders the rendering of the adversarial texture in visible modality. To address these challenges, we propose a shape-shared stacking strategy, where both the visible and infrared patches adopt the same attack shape. This design not only effectively mitigates interactions between infrared and visible patches in the real world but also enhances the attack shapes of visible patches, thereby improving overall attack performance.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets and Evaluation Metrics

We conduct experiments on RGBT234 [12] and LasHeR [13] datasets and assess the effectiveness of our ACAttack by evaluating precision rate (PR) and success rate (SR), both of which are commonly used metrics in tracking tasks. Taking PR as an example, we calculate the Euclidean distance of the center between the predicted bounding box and ground truth box in both RGB and thermal modalities, using the smaller distance to represent the precision RGBT234 provides 234 pairs of RGB and thermal video, with a total frame of about 234K and a maximum of 8K per sequence. LasHeR is comprised of 1224 visible and thermal video pairs, totaling over 730K frame pairs. Since the tracking performance on the background is not of interest, LasHeR performs strict alignment of the object area, allowing the object to share the same ground truth of the bounding box in both visible and thermal modalities. Therefore, we use PR and SR as evaluation metrics.

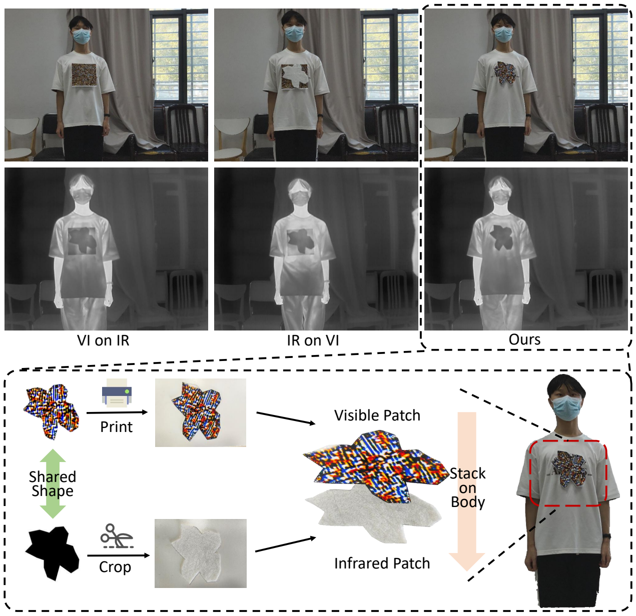
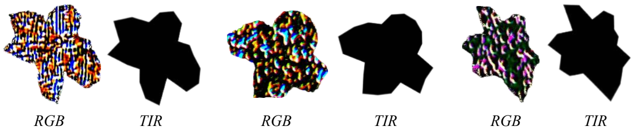


Figure 3. Process of physical implementation.

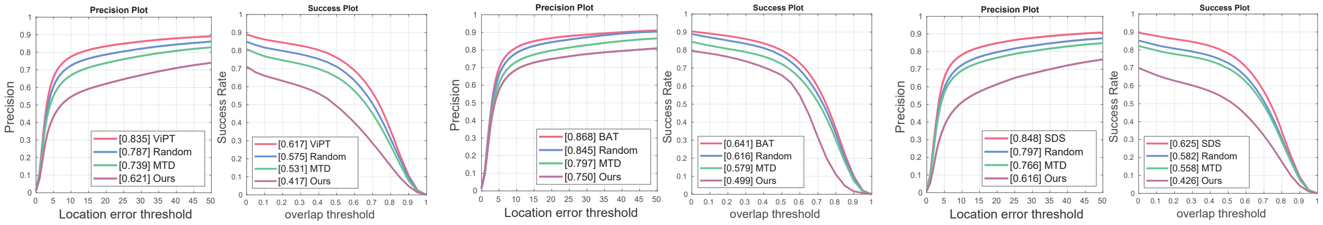


(a) ViPT patch (b) BAT patch (c) SDSTrack patch

Figure 4. Visualization of generated patches.

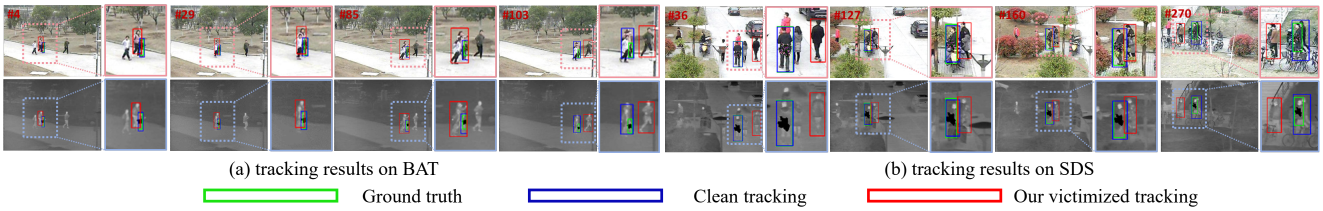
4.1.2. Victimized Trackers and Comparison Attackers

We select several state-of-the-art trackers as targets for our attack, including ViPT [29], BAT [3], and SDSTrack [9]. To demonstrate the challenges in exploiting vulnerabilities in RGB-T trackers, we use a patch composed of random noise as a baseline for comparison, emphasizing the need for meticulous exploration. Furthermore, we compare the performance of our proposed ACAttack with the representative attack method MTD [8], which is specifically designed



(a) ViPT on RGBT234 (b) BAT on RGBT234 (c) SDS on RGBT234

Figure 5. Quantitative comparison of tracking performance on the RGBT234 dataset. The tracking performance of ViPT, BAT, and SDSTrack trackers is reported, including the original performance without attacks and the performance under attacks. **Lower tracking metrics PR and SR represent better attack. Please zoom in for a better view.**



(a) tracking results on BAT (b) tracking results on SDS

🟩 Ground truth 🟦 Clean tracking 🟥 Our victimized tracking

Figure 6. Qualitative comparison of tracking performance on the RGBT234 dataset.

for RGB trackers, highlighting the advantages of our approach in the multi-modal setting.

4.1.3. Implementation Details

The multi-spectral video in the physical domain is captured by a DJI Mavic 3T UAV equipped with thermal and RGB cameras, and the video frame rate is 30 fps. The hyper-parameters in adaptive iteration ξ and ζ is 9 and 0.02. Training epoch in stage1 is set as $M_{stage1} = 180$. Experiments are conducted on the RTX 3090 GPU with PyTorch.

4.2. Comparisons in the Digital Domain

We first validate the attack effectiveness of our ACAttack in the digital domain. It is important to note that we only train on the RGBT234 dataset and generate multi-modal patches $\{p_{vi}, p_{ir}\}$. As shown in Fig. 4, the RGB patch exhibits color and texture, while the thermal patch has an irregular shape, which aligns with the imaging characteristics of each modality. Subsequently, the patches $\{p_{vi}, p_{ir}\}$ generated on RGBT234 are directly applied to the LasHeR dataset to verify their generalization.

4.2.1. Quantitative Evaluation

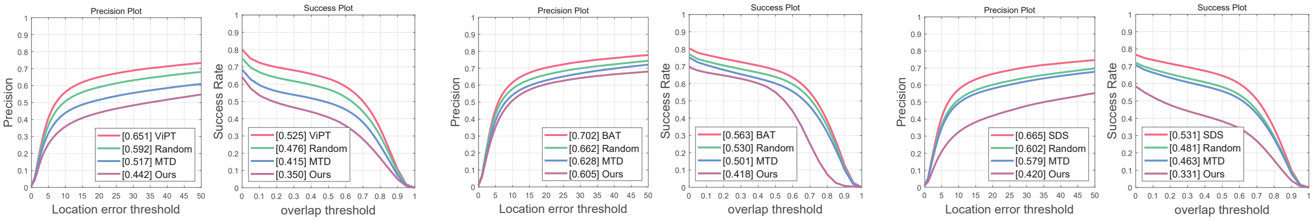
Fig. 5 illustrates a quantitative comparison of the RGBT234 dataset. The results clearly show that, under our attack, the tracking performance of existing state-of-the-art trackers suffers a significant degradation compared to clean tracking conditions. In contrast, random noise only leads to a modest decline in PR and SR, emphasizing that exploiting tracker vulnerabilities goes beyond the simplicity of random noise—it requires a more sophisticated, optimized approach. Additionally, the performance drop observed with MTD is smaller than that of our ACAttack, suggesting that attack methods designed specifically for RGB trackers may not effectively mitigate the feature enhancement resulting from RGB-T coupling. On the other hand, our ACAttack achieves substantial attack success. For instance, against ViPT, ACAttack reduces PR from 0.835 to 0.621 and SR from 0.617 to 0.417. Similarly, for SDSTrack, it lowers PR from 0.848 to 0.616 and SR from 0.625 to 0.426. The substantial performance drops suggest that our ACAttack succeeds in keeping the predicted bounding box far away from actual object, which will be further confirmed in subsequent qualitative results.

4.2.2. Qualitative Evaluation

As shown in Fig. 6, we present the tracking results of BAT and SDSTrack. The clean trackers perform exceptionally well in maintaining precise tracking, while our attack leads to a significant decline in tracking performance. This degradation can be attributed to our progressive generation framework, which iteratively weakens the tracker's deep semantic attention on modalities with high commonality by decoupling multi-modal responses.

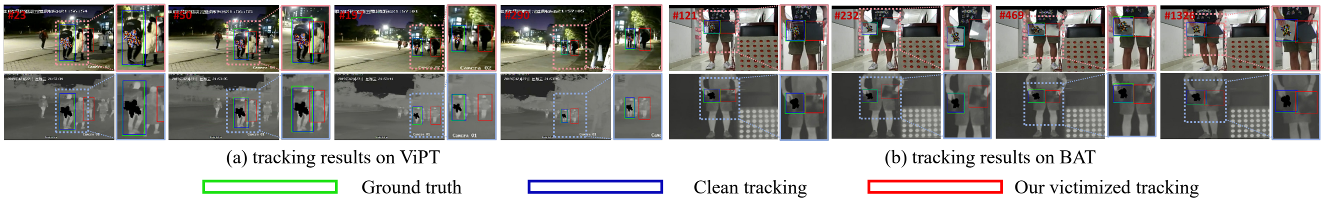
4.3. Generalization Evaluation

We conduct generalization experiments on the LasHeR dataset, with quantitative and qualitative results shown in Fig. 7 and Fig. 8, respectively. Compared to random noise and MTD, our ACAttack leads to a significant drop in tracking performance across all trackers, even without training on LasHeR. Additionally, we present the IoU plots for both



(a) ViPT on LasHeR (b) BAT on LasHeR (c) SDSTrack on LasHeR

Figure 7. Quantitative comparison of tracking performance on the LasHeR dataset. The tracking performance of ViPT, BAT, and SDSTrack trackers is reported, including the original performance without attacks and the performance under attacks. **Lower tracking metrics PR and SR represent better attack. Please zoom in for a better view.**



(a) tracking results on ViPT (b) tracking results on BAT

Figure 8. Qualitative comparison of tracking performance on the LasHeR dataset.

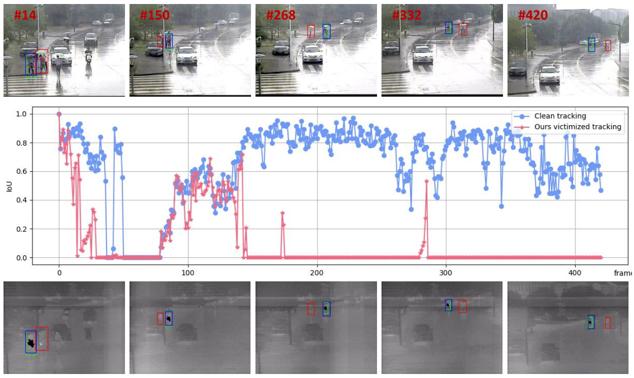


Figure 9. Qualitative comparison of tracking performance on the LasHeR dataset. The blue and red lines represent the IoU variation over frames of the predicted boxes under the clean trackers and the victimized trackers, respectively.

clean and attacked tracking results, as shown in Fig. 9. It is clear that our ACAttack can maintain a sustained attack over extended periods. Due to the existence of our adaptive attack strategy and the modal balance interference loss, the response value of the tracker for the real target is reduced, and then the tracker is easy to deviate from the original target and is attracted by similar targets.

4.4. Application in the Physical Domain

After having verified our adversarial patches in digital scenes, we also extend experiments to demonstrate their efficacy in the physical domain. We directly apply the patches trained in the digital domain to the real world and use aerogel and paper to make thermal and RGB patches for deployment on pedestrians, respectively. A dual-spectral camera in DJI Mavic 3T is used for video capture. Thirty sets of videos of different scenes are taken as test samples. The orientation results of the test are shown in Fig. 10. It can be seen that the tracking prediction bounding box is enlarged and cannot be accurately positioned due to the interference of the multi-modal adversarial patch. Specifically, the optimization of spatio-temporal joint loss makes the patch learn the effect of expanding the tracker's prediction box. Therefore, in the physical world, the tracker will not be able to accurately locate the target after being affected by the adversarial patch.

4.5. Ablation Studies

We conduct ablation studies to assess the effectiveness of our unique design and parameter configuration, including: (I) loss function, (II) parameter K, (III) iteration mode, and (IV) applied modal. The ablation studies are performed on the RGBT234 dataset against ViPT, with quantitative results presented in Table 1.

4.5.1. Loss Function

The loss L_{st} interferes with the tracker from both temporal and spatial dimensions, while L_{mi} is used to disrupt the tracker's semantic perception. To demonstrate their effectiveness, we remove each of them individually, with the results shown in Table 1. In the absence of L_{st} or L_{mi}, the attack performance weakens, demonstrating their role in diminishing the enhanced target localization accuracy achieved through multi-modal interaction.

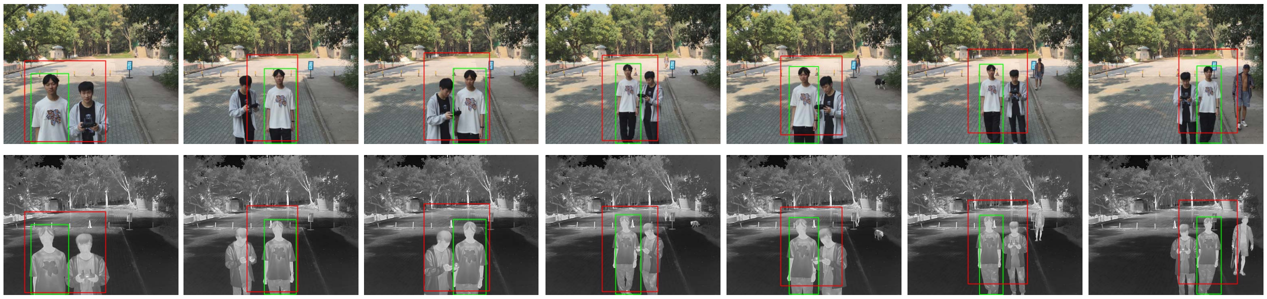


Figure 10. Practical application in the physical domain.

Metric	ViPT	Config. I: loss function		Config. II: parameter K		Config. III: iteration mode		Config. IV: applied modal		Ours
		w/o L_{st}	w/o L_{mi}	K $= 0$	K $= 9$	cross	combine	Only RGB	Only TIR	
PR	0.835	0.709	0.735	0.672	0.651	0.645	0.703	0.691	0.669	0.621
SR	0.617	0.486	0.505	0.450	0.425	0.428	0.482	0.482	0.462	0.417

Table 1. Quantitative comparison of ablation studies, which is performed on the RGBT234 dataset against the ViPT tracker.

4.5.2. Parameter K

In our progressive attack framework, we first employ projected gradient descent to identify K sets of coarse adversarial examples with effective attack performance. In order to verify its effectiveness, we set the number of coarse adversarial samples K growth from 0 to 9 and 18. As shown in the Table 1, as K increases from 0 to 9 and 18, the tracker's PR and SR consistently decrease. This indicates that such coarse-grained adversarial examples can effectively narrow the search space for refined attacks, thus facilitating a more effective attack. Specifically, this progressive method for finding adversarial examples prioritizes identifying multiple sets of coarse adversarial representations from a broad spectrum of noise. Subsequently, multi-modal patch generation refines the adversarial details to produce the final adversarial patch, leveraging numerous samples that contain adversarial information. Consequently, this approach results in an enhancement in performance.

4.5.3. Iteration Mode

One of the key contributions of this paper is the adaptive iterative strategy for attacking the RGB-T tracker. To demonstrate the effectiveness of the adaptive strategy, we conduct ablation experiments using the iterative strategy. The alternating iteration strategy and the joint optimization strategy are selected for the comparison test. The former alternately optimizes the adversarial texture network and the adversarial shape network, while the latter simultaneously propagates the gradient flow to both networks. As shown in Table 1, our adaptive iteration approach can more effectively identify model vulnerabilities and generate more aggressive adversarial patches. Specifically, according to the contribution degree, our strategy can weaken deep semantic attention and break the balance of modality in tracker.

4.5.4. Applied Modal

In order to verify the multi-modal patch joint and single-modal patch attack performance, we try to conduct patch apply modal ablation experiment. Multi-modal patches $\{p_{vi}, p_{ir}\}$ are generated to simultaneously disrupt both RGB and thermal modalities. As shown in the Table 1, we use only one of these patches in an ablation setup. The adversarial patch of a single modal produces a certain attack effect and makes the tracker confused. Evidently, our multi-modal patch achieves the best attack performance, underscoring the necessity of designing joint multi-modal attacks for RGB-T trackers.

5. Conclusion

In this work, we present a pioneering framework for adversarial attacks on RGB-T multi-modal trackers by introducing an adaptive cross-attack mechanism through multi-modal response decoupling. Our approach leverages a modal-aware adaptive attack strategy and introduces novel modal disturbance loss and spatio-temporal joint attack loss to progressively impair the tracker's capability to perceive the target. The shared adversarial shape design also enhances our method's practicality, allowing seamless deployment of multi-modal patches in the real world. Experiments across digital and physical domains confirm the robustness and effectiveness of our approach in evading RGB-T trackers, highlighting the potential and significance of adaptive, multi-modal adversarial attacks in advancing the understanding of tracker vulnerabilities.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62276192).

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 850–865, 2016. 1

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6182–6191, 2019. 2

[3] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bidirectional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 927–935, 2024. 2, 5

[4] Xuesong Chen, Xiyu Yan, Feng Zheng, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Rongrong Ji. One-shot adversarial attacks on visual tracking with dual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10176–10185, 2020. 2

[5] Xuesong Chen, Xiyu Yan, Feng Zheng, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Rongrong Ji. One-shot adversarial attacks on visual tracking with dual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10176–10185, 2020. 2

[6] Xuesong Chen, Canmiao Fu, Feng Zheng, Yong Zhao, Hongsheng Li, Ping Luo, and Guo-Jun Qi. A unified multi-scenario attacking network for visual object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1097–1104, 2021. 1

[7] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2020. 2

[8] Li Ding, Yongwei Wang, Kaiwen Yuan, Minyang Jiang, Ping Wang, Hua Huang, and Z Jane Wang. Towards universal physical attacks on single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1236–1245, 2021. 2, 5

[9] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26551–26561, 2024. 2, 5

[10] Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In *Proceedings of the European Conference on Computer Vision*, pages 69–84, 2020. 1

[11] Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6709–6718, 2021. 1

[12] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 5

[13] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2022. 5

[14] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35:16743–16754, 2022. 2

[15] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2739–2743, 2022. 2

[16] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing*, 30:5613–5625, 2021. 1

[17] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4118–4131, 2025. 2

[18] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016. 2

[19] Wuqiang Qi, Zhuoqun Zhang, and Zhishe Wang. Dmfuse: Diffusion model guided cross-attention learning for infrared and visible image fusion. *Chinese Journal of Information Fusion*, 1(3):226–241, 2024. 1

[20] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Deep learning-based image fusion: A survey. *Journal of Image and Graphics*, 28(1):3–36, 2023. 1

[21] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, Wei Liu, and Houqiang Li. Unsupervised deep representation learning for real-time tracking. *International Journal of Computer Vision*, 129(2):400–418, 2021. 2

[22] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2831–2838, 2022. 5

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 4

[24] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37, 2024. 1

[25] Bin Yan, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 990–999, 2020. 2, 4

[26] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10448–10457, 2021. 1

[27] Fan Zhang, Hanwei Peng, Lingli Yu, Yuqian Zhao, and Baifan Chen. Dual-modality space-time memory network for rgbt tracking. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 2

[28] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5404–5413, 2023. 2

[29] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 2, 5