Cross-Modal Stealth: A Coarse-to-Fine Attack Framework for RGB-T Tracker

Xinyu Xiang[1]*, Qinglong Yan[1]*, Hao Zhang[1]†, Jianfeng Ding[2], Han Xu[3],
Zhongyuan Wang[2], Jiayi Ma[1]†

[1]Electronic Information School, Wuhan University, Wuhan 430072, China,
[2]School of Computer Science, Wuhan University, Wuhan 430072, China,
[3]School of Automation, Southeast University, Nanjing 210096, China
xiangxinyu@whu.edu.cn, qinglong_yan@whu.edu.cn, zhpersonalbox@gmail.com, jyma2010@gmail.com

Abstract

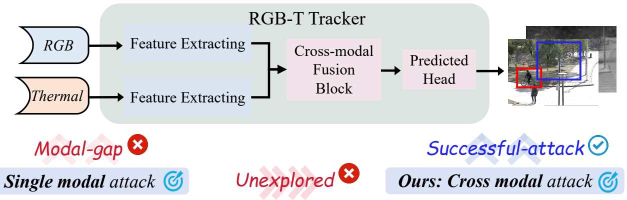
Current research on adversarial attacks mainly focuses on RGB trackers, with no existing methods for attacking RGB-T cross-modal trackers. To fill this gap and overcome its challenges, we propose a progressive adversarial patch generation framework and achieve cross-modal stealth. On the one hand, we design a coarse-to-fine architecture grounded in the latent space to progressively and precisely uncover the vulnerabilities of RGB-T trackers. On the other hand, we introduce a correlation-breaking loss that disrupts the modal coupling within trackers, spanning from the pixel to the semantic level. These two design elements ensure that the proposed method can overcome the obstacles posed by cross-modal information complementarity in implementing attacks. Furthermore, to enhance the reliable application of the adversarial patches in real world, we develop a point tracking-based reprojection strategy that effectively mitigates performance degradation caused by multi-angle distortion during imaging. Extensive experiments demonstrate the superiority of our method. Our code is provided at https://github.com/Xinyu-Xiang/CMS.

Introduction

Visual Object Tracking (VOT) aims to recognize and locate a tracked object in subsequent frames, starting from its initial appearance in the first frame of a visible video. Due to the representation limitations of the RGB modality in harsh conditions (Liu et al. 2024), researchers prefer to combines complementary information from RGB and thermal modalities to achieves more accurate and environmentally robust tracking (Zhu et al. 2023; Liu et al. 2023). Tracking finds extensive application in critical domains including autonomous driving and intelligent surveillance (Zhang et al. 2022, 2023), and addressing its security challenges is paramount. Fortunately, adversarial attacks aid in exploring potential vulnerabilities comprehensively and serve as a critical mechanism to bolster model security, which has attracted a great deal of scholarly attention.

In the past decades, numerous tracker attack techniques have been proposed, including one-stage attack methods and two-stage attack methods. The one-stage attack methods (Zhao et al. 2023; Lou et al. 2023; Li et al. 2021) directly employ a network or a perturbation generation strategy to produce adversarial examples. Although effective attack performance can be achieved, it is challenging to thoroughly examine the model vulnerabilities. Different from one-stage attack, two-stage attack methods (Jia et al. 2021; Li et al. 2022b) can exploit more hidden model defects through the use of coarse attacks in the initial phase and fine-tuning in subsequent stages. Specifically, Stage I extensively identifies potential model vulnerabilities to attain a preliminary level of attack. Furthermore, Stage II employs a more refined attack strategy to target the critical vulnerabilities of the model accurately. Consequently, in comparison with one-stage attack, the two-stage methods can more meticulously uncover the security issues within the model.

Although existing adversarial attack methods (Yan et al. 2020; Huang et al. 2024) have demonstrated the capability to fool trackers to some extent, there are still some challenges that need to be addressed. Firstly, as shown in Fig. 1,



(a) Attack RGB Tracker

(b) Attack RGB-T Tracker

Attack method	Attack modal	Attack domain
PAT (ICCV2019)	RGB	digital
CSA (CVPR2020)	RGB	digital
TTAttack (ECCVW2022)	RGB	digital
Pluggable (TIFS2023)	RGB	digital
Context-Guided (TMM2024)	RGB	digital
CMS (Ours)	RGB & Thermal	digital & physical

Figure 1: The superiority of our attack model: *(a)* Current attack methods for RGB-T trackers *remain unexplored*, and existing attack methods targeting RGB trackers achieve a limited attack effectiveness due to *modal gap*, *i.e.,* failing to disrupt the feature enhancement resulting from multi-modal coupling. *(b)* We pioneer a coarse-to-fine attack framework tailored for RGB-T tracking, combined with a robust physical attack strategy, demonstrating strong and versatile capabilities in *both digital and physical domains*.

*These authors contributed equally.

†Corresponding authors.

current research focuses primarily on attacks against RGB trackers, with methods for RGB-T tracking remaining unexplored. This gap poses a significant threat to the robustness and security of RGB-T tracking systems in practical applications. Therefore, how to create an adversarial attack framework for RGB-T trackers is the first challenge. Secondly, attacking RGB-T trackers presents inherent technical difficulties. On the one hand, the cross-modal trackers are capable of compensating for the disturbed part of a single modal using modal complementary information. On the other hand, it is challenging for adversarial perturbations to obfuscate the correlation between the cross-modal template image and search images. Hence, how to gradually dig and break the modal coupling between RGB and thermal modality is the second challenge. Moreover, the challenge of transitioning attack methods from the digital domain to practical physical applications constrains the progress of adversarial attack techniques. When capturing the patch in the real world, different shooting angles of the sensor may lead to distortion of the patch in the photo, which will degrade the attack performance. Despite certain pioneering explorations in physical attacks, existing work (Ding et al. 2021) mainly focuses on applying random transformations, which does not align with the true physical significance. Thus, how to guarantee multi-angle robustness in physical domain is the third challenge.

To address the aforementioned challenges, we propose a coarse-to-fine attack framework for RGB-T trackers. To our knowledge, our work is the first attempt to attack cross-modal tracking methods. Firstly, we devise a coarse-to-fine network to gradually break the coupling of two modalities and progressively generate the thermal patch with attack shape and the RGB patch with attack texture from the same latent space. Specifically, the RGB patch with adversarial texture is first generated via latent code, and subsequently, the RGB patch and the optimized latent code are modulated by two distinct networks to produce the enhanced RGB patch and the thermal patch with adversarial structures, respectively. Secondly, we design two loss functions from pixel-level to semantic-level, which are capable of interfering with image block correlation and disrupting RGB-T feature response, respectively. Finally, to mitigate the impact of angle changes on the image appearance of adversarial samples, we propose a patch reprojection strategy based on point tracking. This strategy leverages the consistency of key point correspondences to dynamically adjust the patch's position, thus ensuring effective attack performance under multiple imaging angles in physical domain.

The contributions of this paper are as follows:

- We propose a cross-modal feasible adversarial patch to misguide RGB-T trackers. To the best of our knowledge, it is the first work to evade RGB-T trackers.
- We introduce a coarse-to-fine integrated cross-modal patch generation strategy from both architecture and loss function, to destroy pixel-level and semantic-level modal coupling between RGB and thermal modality.
- We design a point tracking-based patch reprojection strategy, so that cross-modal patches can be better extended from the digital to the physical world.
- Experimental results show that cross-modal patches can efficiently fool RGB-T trackers in standard RGB-T tracking datasets and real scenes.

Related Work

Visual Object Tracking

Numerous RGB tracking methods have been proposed and achieved impressive tracking performance (Nam and Han 2016; Bhat et al. 2019; Danelljan, Gool, and Timofte 2020; Bertinetto et al. 2016). However, the unimodal information provision limits robustness under harsh conditions, such as occlusion, low illumination, *etc* (Yi et al. 2024). To address this concern, researchers propose the RGB-T tracking paradigm, which draws on the strengths of both modalities to achieve robust tracking. APFNet (Xiao et al. 2022) proposes an attribute-based progressive fusion network with specific branches for fusing different attributes. ViPT (Zhu et al. 2023) introduces a prompt learning paradigm to convert thermal modalities into visual prompt, motivating the RGB-based foundation model to adapt downstream RGB-T tracking. Further, BAT (Cao et al. 2024) proposes a general bi-directional adapter that does not preset the dominant modality, enabling multiple modalities to cross-prompt.

Adversarial Attacks

The primary aim of adversarial attacks is to scrutinize the inherent vulnerabilities of deep learning models and their susceptibility to malicious exploitation. Current attack strategies in the tracking domain predominantly target RGB trackers. For instance, CSA (Yan et al. 2020) introduces a cooling-contraction strategy that cools hotspot regions in heatmaps where objects are detected, thereby directly contracting the predicted bounding box. APYVOT (Chen et al. 2020) proposes an optimization objective function with a dual attention mechanism to generate adversarial perturbations, which disrupt the tracking by interfering with the initial frame only. However, these methods are unable to interfere with the salient feature enhancement brought by RGB-T multi-modal interactions, rendering them less effective against RGB-T trackers. Thus, developing a specialized adversarial attack algorithm for RGB-T tracking is crucial.

Methodology

Problem Formulation

In RGB-T tracking, the tracker locates object within the search images based on a given template. Consequently, the core of tracking can be distilled into the challenge of semantic similarity learning. Furthermore, the joint participation of RGB and thermal modalities integrates complementary information and promotes more robust tracking. Specifically, the RGB image provides rich texture details, while the thermal image captures the thermal radiation information of object, which is reflected by high intensity values. Therefore, with respect to *i) the semantic similarity measure and ii) modal properties of RGB-T data*, we design a coarse-to-fine progressive generation framework to attack the RGB-T

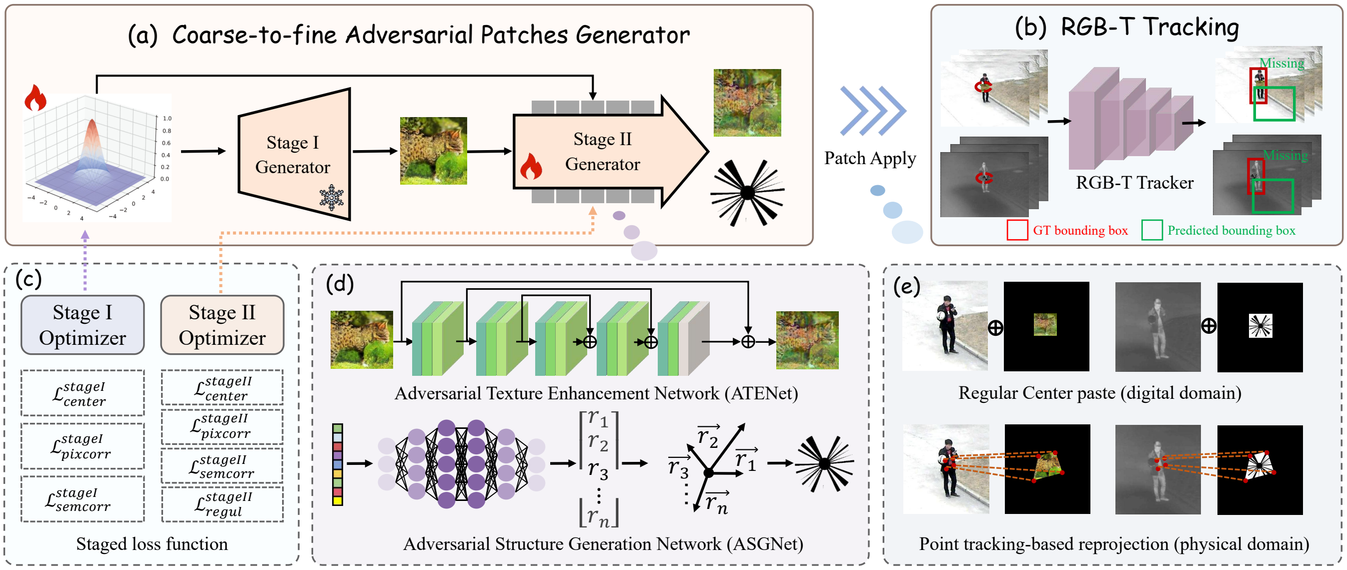

(a) Coarse-to-fine Adversarial Patches Generator
Stage I Generator
Stage II Generator
Patch Apply
(b) RGB-T Tracking
Missing
RGB-T Tracker
Missing
GT bounding box
Predicted bounding box
(c)
Stage I Optimizer
Stage II Optimizer
$\mathcal{L}^{stageI}_{center}$
$\mathcal{L}^{stageI}_{pixcorr}$
$\mathcal{L}^{stageI}_{semcorr}$
$\mathcal{L}^{stageII}_{center}$
$\mathcal{L}^{stageII}_{pixcorr}$
$\mathcal{L}^{stageII}_{semcorr}$
$\mathcal{L}^{stageII}_{regul}$
Staged loss function
(d)
Adversarial Texture Enhancement Network (ATENet)
r_1 r_2 r_3 r_n
$\overrightarrow{r_2}$ $\overrightarrow{r_1}$ $\overrightarrow{r_3}$ $\overrightarrow{r_n}$
Adversarial Structure Generation Network (ASGNet)
(e)
Regular Center paste (digital domain)
Point tracking-based reprojection (physical domain)

Figure 2: Overview of the proposed attack pipeline CMS, which consists of a two-stage generative framework from coarse to fine. Moreover, a reprojection strategy based on point tracking is proposed to guarantee robust attacks in the physical domain.

trackers and achieve cross-modal stealth (**CMS**). The overall workflow of our CMS is presented in Fig. 2, which is categorized into Stage I and Stage II.

Stage I Considering the core issue in tracking, *i.e.,* semantic similarity metric, we design an adversarial semantic optimization network ($ASONet$) that seeks an optimal adversarial vector l_{adv} through ***semantic-aware adversarial latent optimization***. This optimized l_{adv} is utilized to generate the first-stage RGB patch δ^{stageI}_{rgb}. Remarkably, $ASONet$ is constructed using the generator from StyleGAN2 (Karras et al. 2020), which is pre-trained on real-world images, boasting exceptional generative capabilities and a nuanced understanding of semantic representations. Consequently, the patch δ^{stageI}_{rgb} encapsulates realistic adversarial semantics, effectively disrupting the semantic similarity between the template and the object in search images.

Stage II Building on the adversarial semantic disruptions of Stage I, we introduce a ***coupled adversarial texture-structure generation*** strategy that targets the different characteristics of RGB-thermal data. This strategy aims to undermine the effective information enhancement gained from complementary feature coupling, enabling a finer attack. For the RGB modality, which features clear textures, we focus on embedding adversarial textures. Specifically, we feed the first-stage RGB patch δ^{stageI}_{rgb} into an adversarial texture enhancement network ($ATENet$) to generate δ_{rgb}, refining the attack by layering adversarial textures onto the semantic disruptions. For the thermal modality reflecting temperature distribution through pixel intensity, we input the adversarial vector l_{adv} into an adversarial structure generation network ($ASGNet$). $ASGNet$ produces the thermal patch δ_{tir}, a dynamically shaped black mask that implements the attack through adversarial structural generation.

With the coarse-to-fine adversarial optimization of Stage I and Stage II, the generated multi-modal adversarial patches $\{\delta_{rgb}, \delta_{tir}\}$ demonstrate superior attack performance in the digital domain. However, it is important to note that during this digital optimization process, the patches are consistently applied in a square "🟩" format across all frames. However, the orientation and viewpoint of object in the physical world are constantly changing, which makes the shape of patches also vary, posing a serious challenge to the attack robustness. Thus, an angle-robust strategy is desired in digital domain training, with the added challenge that these angle variations must correspond to actual physical significance, rather than mere random transformations. To achieve the ***Angle-Robust Physical Domain Guarantee***, we design a point tracking-based patch reprojection strategy that integrates the homography transformation with keypoint tracking, acquiring object affixed with patch under consistent angles. With this reprojection strategy, our CMS exhibits more robust attack performance in the physical world.

Next, we first specifically describe the semantic-aware adversarial latent optimization in Stage I and coupled adversarial texture-structure generation in Stage II, to achieve digital domain attacks. Subsequently, we introduce the point tracking-based reprojection strategy to facilitate robust physical domain application.

Semantic-Aware Adversarial Latent Optimization

Previous works optimize adversarial patches directly in pixel space, often leading to unnatural and conspicuous results. Drawing inspiration from NatPatch (Hu et al. 2021), we optimize a vector in the latent space using the adversarial semantic optimization network ($ASONet$). The latent vector $l \in \mathbb{R}^d$ is randomly sampled from a standard normal distribution. As shown in Fig. 2(a), we feed l into the $ASONet$, and generate the patch δ^{stageI}_{rgb} for Stage I, formulated as:

$$\delta^{stageI}_{rgb} = ASONet(l). \tag{1}$$

It is worth noting that we only optimize the latent vector l to get l_{adv}, rather than optimizing the generator $ASONet$. Leveraging the robust semantic representation capabilities of the pre-trained StyleGAN2, the δ^{stageI}_{rgb} generated from l_{adv} excellently integrates both naturalistic and adversarial semantics. Subsequently, we paste δ^{stageI}_{rgb} to the center of object in the original images, formulated as:

$$I^{(rgb,s)}_{\delta^{stageI}_{rgb}} = I^{(rgb,s)} \odot (1 - M^{s}_{rgb}) + \delta^{stageI}_{rgb} \odot M^{s}_{rgb}, \tag{2}$$

where $I^{(rgb,s)}$ denotes the original RGB image in s-th frame. M^{s}_{rgb} represents the binary mask for applying adversarial patch. In the digital domain, M^{s}_{rgb} is generated based on the annotated ground truth of bounding box $\mathcal{B}box_{gt}$. In the physical domain, a physically consistent reprojection strategy based on point tracking is designed to generate M^{s}_{rgb}. Notably, the patch is added when $s \geq 2$, and for the initial frame ($s = 1$), we have $I^{(rgb,1)}_{\delta^{stageI}_{rgb}} = I^{(rgb,1)}$. Then, we feed $I^{(rgb,s)}_{\delta^{stageI}_{rgb}}$ and $I^{(tir,s)}$ to the RGB-T tracker $T(\cdot)$ and predict the bounding box, formulated as:

$$\mathcal{B}box^{stageI}_{pred} = T(I^{(rgb,s)}_{\delta^{stageI}_{rgb}}, I^{(tir,s)}), \tag{3}$$

where $I^{(tir,s)}$ is the s-th original thermal image, and $\mathcal{B}box^{stageI}_{pred}$ is the predicted bounding box in Stage I.

To achieve attack, we shift the center of $\mathcal{B}box^{stageI}_{pred}$ away from center of $\mathcal{B}box_{gt}$, with the following center-away loss:

$$\mathcal{L}^{stageI}_{center} = -\|(C(\mathcal{B}box^{stageI}_{pred}) - C(\mathcal{B}box_{gt}))\|^2_2, \tag{4}$$

where $C(\cdot)$ is the center coordinate of bounding box. We further introduce two correlated interference losses to disrupt the tracker's performance at both the pixel and semantic levels, which are represented as follows:

$$\begin{aligned}\mathcal{L}^{stageI}_{pixcorr} = -\|&\arg\max_{x,y}(Corr(z^{rgb}, x^{(rgb,s)})) \\ &-\arg\max_{x,y}(Corr(z^{rgb}, x^{(rgb,s)}_{\delta^{stageI}_{rgb}}))\|^2_2,\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}^{stageI}_{semcorr} = -\|&\arg\max_{x,y}(R(z^{rgb}, z^{tir}, x^{(rgb,s)}, x^{(tir,s)})) \\ &-\arg\max_{x,y}(R(z^{rgb}, z^{tir}, x^{(rgb,s)}_{\delta^{stageI}_{rgb}}, x^{(tir,s)}))\|^2_2.\end{aligned} \tag{6}$$

Let $m \in \{rgb, tir\}$ represent different modalities of RGB and thermal, $z^{m} \in \mathbb{R}^{w_z \times h_z \times c}$ is the template image cropped from $I^{(m,1)}$, and $x^{(m,s)} \in \mathbb{R}^{w_x \times h_x \times c}$ is the search image cropped from $I^{(m,s)}$ ($s \geq 2$). Similarly, the adversarial search image $x^{(rgb,s)}_{\delta^{stageI}_{rgb}}$ is from $I^{(rgb,s)}_{\delta^{stageI}_{rgb}}$. $Corr(\cdot)$ is the correlation operator, $R(\cdot)$ is the feature response map of the tracker, and (x, y) is coordinate on the correlation or response map. Specifically, $\mathcal{L}^{stageI}_{pixcorr}$ affects the RGB-T tracker by increasing the disparity between maximum correlation points of the template on clean and adversarial search images. Since the RGB-T tracker couples the complementary information of the two modalities through multi-modal interaction, its feature response map can reflect a clearer and more discriminative distinction between the object and background. Therefore, we use $\mathcal{L}^{stageI}_{semcorr}$ to disrupt the multi-modal coupling from a semantic perspective.

Finally, the full loss function of Stage I is defined as:

$$\mathcal{L}_{stageI} = \alpha_1 \cdot \mathcal{L}^{stageI}_{center} + \alpha_2 \cdot \mathcal{L}^{stageI}_{pixcorr} + \alpha_3 \cdot \mathcal{L}^{stageI}_{semcorr}, \tag{7}$$

where α_1, α_2 and α_3 are the hyper-parameters for balancing each sub-loss of Stage I.

Coupled Adversarial Texture-Structure Generation

Considering the characteristics of RGB and thermal modalities, we propose a coupled adversarial texture-structure generation strategy in Stage II, to generate multi-modal patches. As shown in Fig. 2(d), we feed δ^{stageI}_{rgb} to the adversarial texture enhancement network ($ATENet$), formulated as:

$$\delta_{rgb} = ATENet(\delta^{stageI}_{rgb}), \tag{8}$$

where δ_{rgb} is the final RGB patch with both adversarial semantics and rich adversarial textures, and $ATENet$ consists of residual connections and a series of convolutional layer - activation function - batch normalization.

For thermal patch, we design the adversarial structure generation network ($ASGNet$), which takes a latent code as input and uses MLPs to generate the lengths of a number of pole vectors. Through optimizing MLPs, these varying lengths construct the shape of the black patch. The above process is formulated as:

$$\delta_{tir} = ASGNet(l_{adv}), \tag{9}$$

where δ_{tir} refers to the generated thermal patch. It is worth emphasizing that we modulate the latent l_{adv} to generate the adversarial shape. As presented in Fig. 2(e), we apply the multi-modal patches $\{\delta_{rgb}, \delta_{tir}\}$ to original images in a center-pasted manner, and get the adversarial RGB image $I^{(rgb,s)}_{\delta_{rgb}}$ and thermal image $I^{(tir,s)}_{\delta_{tir}}$. They are fed to the RGB-T tracker $T(\cdot)$ to predict the bounding box $\mathcal{B}box^{stageII}_{pred}$. Similarly, we employ the following center-away loss, pixel-level correlated interference loss, and semantic-level correlated interference loss:

$$\mathcal{L}^{stageII}_{center} = -\|(C(\mathcal{B}box^{stageI}_{pred}) - C(\mathcal{B}box_{gt}))\|^2_2, \tag{10}$$

$$\begin{aligned}\mathcal{L}^{stageII}_{pixcorr} = -\sum_{m}\|&\arg\max_{x,y}(Corr(z^{m}, x^{(m,s)})) \\ &-\arg\max_{x,y}(Corr(z^{m}, x^{(m,s)}_{\delta_m}))\|^2_2,\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}^{stageII}_{semcorr} = -\|&\arg\max_{x,y}(R(z^{rgb}, z^{tir}, x^{(rgb,s)}, x^{(tir,s)})) \\ &-\arg\max_{x,y}(R(z^{rgb}, z^{tir}, x^{(rgb,s)}_{\delta_{rgb}}, x^{(tir,s)}_{\delta_{tir}}))\|^2_2.\end{aligned} \tag{12}$$

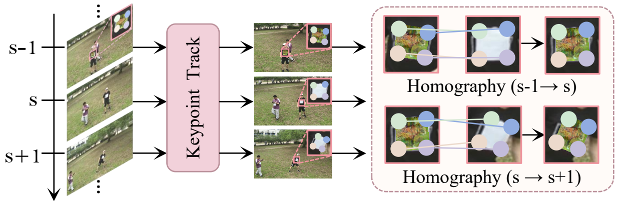


Figure 3: Point tracking-based reprojection strategy.

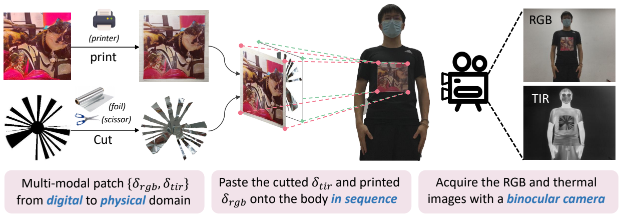

(printer)
print
(foil)
(scissor)
Cut
RGB
TIR
Multi-modal patch $\{\delta_{rgb}, \delta_{tir}\}$ from *digital* to *physical* domain
Paste the cutted δ_{tir} and printed δ_{rgb} onto the body *in sequence*
Acquire the RGB and thermal images with a *binocular camera*

Figure 4: Process of Physical implementation.

In addition, to ensure that the adversarial semantics obtained by the first-stage patch δ_{rgb}^{stageI} are not excessively disruptive, we also incorporate a realism-preserving regularization, which is denoted as:

$$\mathcal{L}_{regul}^{stageII} = \|\delta_{rgb} - \delta_{rgb}^{stageI}\|_2^2. \tag{13}$$

Finally, the full loss function of Stage II is defined as:

$$\begin{aligned} \mathcal{L}_{stageII} = \beta_1 \cdot \mathcal{L}_{center}^{stageII} + \beta_2 \cdot \mathcal{L}_{pixcorr}^{stageII} \\ + \beta_3 \cdot \mathcal{L}_{semcorr}^{stageII} + \beta_4 \cdot \mathcal{L}_{regul}^{stageII}, \end{aligned} \tag{14}$$

where $\beta_1, \beta_2, \beta_3$ and β_4 are employed to control the trade-off between different losses of Stage II.

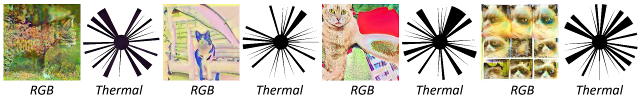
Angle-Robust Physical Domain Guarantee

Point Tracking-based Patch Reprojection Fig. 3 shows the implementation of our reprojection strategy. Specifically, we first capture the coordinates of keypoints frame-by-frame using the top-performing point-tracking algorithm, TAPIR (Doersch et al. 2023). The keypoints in the first frame, *i.e.*, $\{(x_m^{1k}, y_m^{1k})\}_{k=1}^4$, are manually provided to define the initial application range of patch, while the keypoints predicted by TPAIR in the subsequent s-th frame are defined as $\{(x_m^{sk}, y_m^{sk})\}_{k=1}^4$. Across the ($s$-1)-th to s-th frame, we can solve for the homography matrix H_m^{s-1} based on $\{(x_m^{(s-1)k}, y_m^{(s-1)k})\}_{k=1}^4$ and $\{(x_m^{sk}, y_m^{sk})\}_{k=1}^4$. The binary mask M_m^{s-1} is then transformed with H_m^{s-1}, and get M_m^s:

$$M_m^s = H_m^{s-1} M_m^{s-1}. \tag{15}$$

Finally, according to Eq. (2), we can generate $I_{\delta_m}^{(m,s)}$, the adversarial images with patch under various angles and consistent with physical implications.

Physical Implementation After obtaining angle-robust patches $\{\delta_{rgb}, \delta_{tir}\}$, we apply them to the physical world. Specifically, as shown in Fig. 4, we use a color printer to obtain RGB patch δ_{rgb}, and cut thermal patch δ_{tir} from aluminum foil, which possesses high reflectivity and reflects most thermal radiation emitted by the object back to its surface, thereby preventing heat transfer. Next, we sequentially paste the cropped δ_{tir} and the printed δ_{rgb} onto the object. We then use a binocular camera to capture RGB-T videos, which serves as the test set in the physical domain.



(a) ViPT patch (b) BAT patch (c) APFNet patch (d) MANet patch

Figure 5: Visualization of generated patches.

Experiments

Experimental Settings

Datasets and Evaluation Metrics We perform experiments on RGBT234 (Li et al. 2019) and LasHeR (Li et al. 2022a) datasets, and evaluate our attack effectiveness using the precision rate (PR) and success rate (SR), which are the classical metrics on tracking tasks.

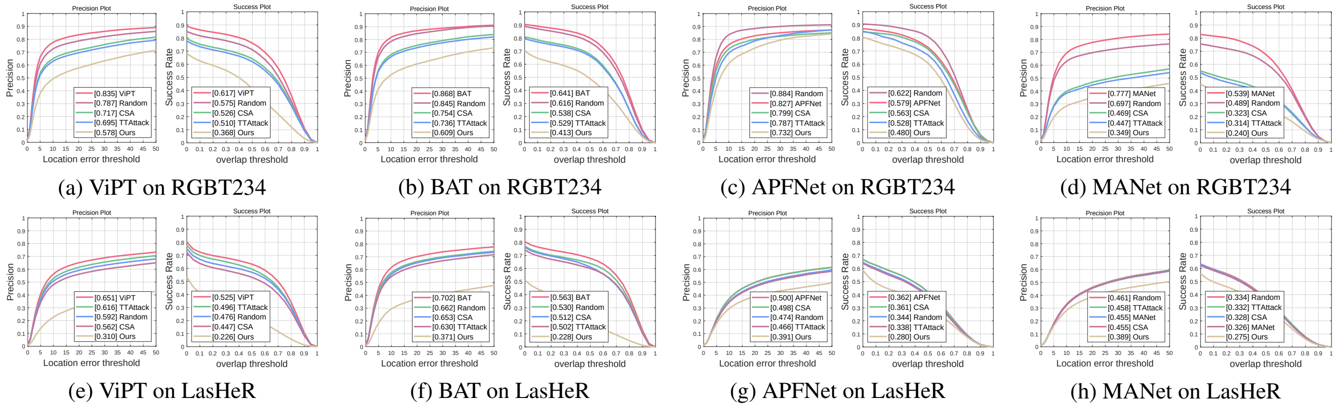
Victimized Trackers and Comparison Attackers We select several SOTA trackers as victims, including ViPT (Zhu et al. 2023), BAT (Cao et al. 2024), APFNet (Xiao et al. 2022), and MANet (Long Li et al. 2019). The first two are powerful offline trackers, while the latter two exemplify typical online trackers. We compare CMS with two representative attack methods, including CSA (Yan et al. 2020) and TTAttack (Nakka and Salzmann 2022). Notably, CSA and TTAttack are designed for RGB trackers, and the patches they generate manifest as perturbations, limiting their applicability in the physical world. Additionally, a patch composed of random noise is also employed as a comparison.

Implementation Details For our two-stage framework, we train Stage I for 80 epochs, and then train Stage II for 30 epochs. The hyper-parameters for balancing each subloss are empirically set as $\alpha_1 = 1.0$, $\alpha_2 = 0.1$, $\alpha_3 = 50.0$, $\beta_1 = 0.1$, $\beta_2 = 0.1$, $\beta_3 = 50.0$, and $\beta_4 = 1.0$. All experiments are conducted on the NVIDIA TITAN RTX GPU with PyTorch.

Comparisons on Digital Domain

We first consider adversarial attack in the digital domain, where our CMS is trained on the RGBT234 dataset and generate the multi-modal patches $\{\delta_{rgb}, \delta_{tir}\}$, as illustrated in Fig. 5. Considering that adversarial patches need to be more natural and realistic, we utilize the StyleGAN2 which is capable of generating real cats to optimize our RGB patches. Hence, the RGB patches carry the semantics of a "cat". The thermal patch consists of shape-variable black mask, reflecting the characteristics of thermal imaging.

Quantitative Evaluation The quantitative comparisons on the RGBT234 dataset are illustrated in Fig. 6(a)-(d), which reveal significant performance degradation in state-of-the-art trackers under our attack. For clearer comparison, specific performance drops are provided in Table 1. Among



(a) ViPT on RGBT234

(b) BAT on RGBT234

(c) APFNet on RGBT234

(d) MANet on RGBT234

(e) ViPT on LasHeR

(f) BAT on LasHeR

(g) APFNet on LasHeR

(h) MANet on LasHeR

Figure 6: Quantitative comparison of tracking performance on the RGBT234 and LasHeR datasets. The tracking performance of ViPT, BAT, APFNet, and MANet 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.

Method	PR on RGBT234				SR on RGBT234				PR on LasHeR				SR on LasHeR			
	Random	**CSA**	**TTAttack**	**Ours**	**Random**	**CSA**	**TTAttack**	**Ours**	**Random**	**CSA**	**TTAttack**	**Ours**	**Random**	**CSA**	**TTAttack**	**Ours**
ViPT (CVPR$_{2023}$)	0.048	0.118	0.140	**0.257**	0.042	0.091	0.107	**0.249**	0.059	0.089	0.035	**0.341**	0.049	0.078	0.029	**0.299**
BAT (AAAI$_{2024}$)	0.023	0.114	0.132	**0.259**	0.025	0.103	0.112	**0.228**	0.040	0.049	0.072	**0.331**	0.033	0.051	0.061	**0.335**
APFNet (AAAI$_{2022}$)	-0.057	0.028	0.040	**0.095**	-0.043	0.016	0.051	**0.099**	0.026	0.002	0.034	**0.109**	0.018	0.001	0.024	**0.082**
MANet (TIP$_{2021}$)	0.080	0.308	0.330	**0.428**	0.050	0.216	0.225	**0.299**	-0.006	0.000	-0.003	**0.066**	-0.008	-0.002	-0.006	**0.051**

Table 1: Quantitative comparison of tracking performance *drop*. A larger drop indicates more effective attack. (Bold: optimal)

Metric	ViPT	Ours	drop
PR	0.998	0.709	0.289
SR	0.835	0.613	0.222

Table 2: Quantitative comparison on the physical domain.

all trackers, random noise only achieves a maximum reduction of 0.080 in PR and 0.050 in SR, highlighting the challenge of identifying tracker vulnerabilities. CSA and TTAttack also show smaller drops than our CMS, demonstrating that RGB-based attacks are less effective against feature enhancement brought by RGB-T coupling. In contrast, our CMS outperforms them significantly. Against the offline tracker ViPT, CMS achieves reductions of 0.257 in PR and 0.249 in SR. Even against the online tracker that performs parameter updates, CMS maintains superior attack, *e.g.,* decreasing the PR and SR of MANet by 0.428 and 0.299. Additionally, we calculate the Euclidean distances between response peaks for CSA, TTAttack, and our CMS compared to the clean map, which are 0.868, 2.071, and 4.381, demonstrating the superior feature interference of our CMS.

Interestingly, we observe a phenomenon also noted in MTD (Ding et al. 2021): tracking performance improves under random noise compared to clean conditions, especially for online trackers like APFNet and MANet. This may be because random noise provides a consistent reference for localization, allowing online trackers to efficiently adjust parameters in response to appearance changes of object.

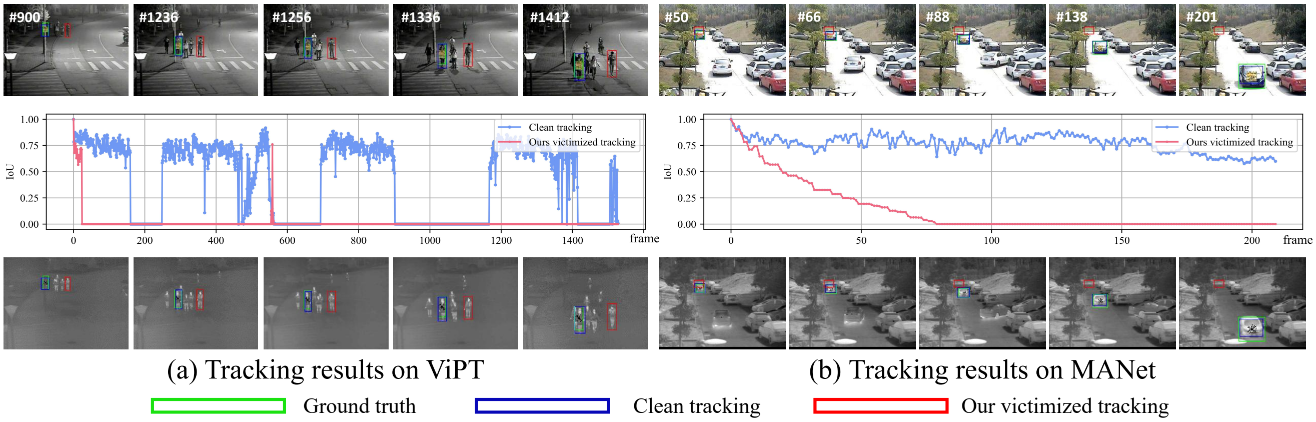
Qualitative Evaluation As shown in Fig. 7, we visualize two groups of tracking results. While the clean trackers excel in maintaining accurate tracking, our attack induces a rapid and substantial decline in tracking performance. The visualizations and IoU trends confirm that attack is executed swiftly and maintained throughout the video. This can be attributed to our coarse-to-fine architecture, which mines adversarial semantics, textures, and structures. Furthermore, the correlated interference losses at both pixel and semantic level effectively break the information coupling, reducing the prominent feature responses of object area.

Generalization Evaluation

Moreover, we conduct generalization experiments on the LasHeR dataset, with results and drops shown in Fig. 6(e)-(h) and Table 1. Among all trackers, our CMS achieves a reduction in PR and SR by at least 0.066 and 0.051, and at most 0.341 and 0.335, significantly outperforming other methods. Additionally, Qualitative results in Fig. 8 further demonstrate its strong generalization capabilities.

Application on Physical Domain

We extend our experiments into the physical domain. We capture training videos using an binocular camera, and apply our reprojection strategy to train the generator and produce adversarial patches. These patches are then tested on physical objects, with results shown in Table 2 and Fig. 9. PR and SR are reduced by 0.289 and 0.222, respectively, while the visualized results also illustrate that our adversarial patches successfully interfere with the tracking, showing the deployability and reliability in real-world applications of our CMS.



(a) Tracking results on ViPT

(b) Tracking results on MANet

Figure 7: Qualitative comparison of tracking performance on the RGBT234 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.

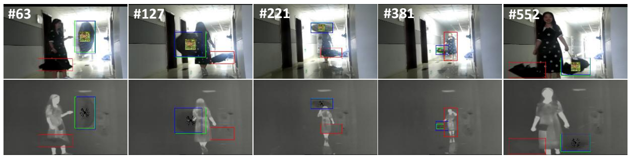


Figure 8: Qualitative comparison on the LasHeR dataset.

Ablation Studies

We perform ablation studies to verify the validity of parameter setting and our specific designs, conducted on RGBT234 dataset against ViPT, with results shown in Fig. 10.

Patch Size Ratios We test different patch size ratios to assess its impact. As shown in Fig. 10(a), the attack performance gradually improves as the ratio increases, which aligns with our expectations. However, considering the feasibility in the physical domain, we select a ratio of 1.6 in our experiments, despite larger ratios yielding stronger attacks.

Two-stage Strategy A two-stage framework from coarse to fine is designed to generate $\{\delta_{rgb}^{stageI}, \delta_{rgb}, \delta_{tir}\}$. As shown in Fig. 10(b), we employ the patch of only one stage as different ablation settings. Obviously, our two-stage patches obtain the optimal attack performance, proving the validity of two-stage strategy that systematically seeks adversarial semantics, textures and structures.

Loss Functions $\mathcal{L}_{pixcorr}$ disrupts the image correlation at pixel level, and $\mathcal{L}_{pixcorr}$ interferes feature responses from semantic level. We remove them separately and obtain results shown in Fig. 10(c). The attack performance diminishes without $\mathcal{L}_{pixcorr}$ or $\mathcal{L}_{semcorr}$, confirming their effect in degrading precise localization.

Reprojection Strategy To address the real challenge of patch deformation, we design a reprojection strategy. We validate its effectiveness by switching to center pasting and random transformation. As shown in Fig. 10(d), our CMS achieves the greatest drop of 0.289 and 0.222 on PR and SR, demonstrating our robustness in the physical domain.

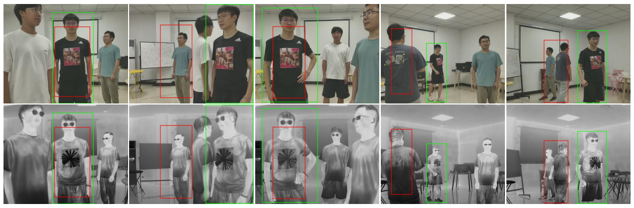
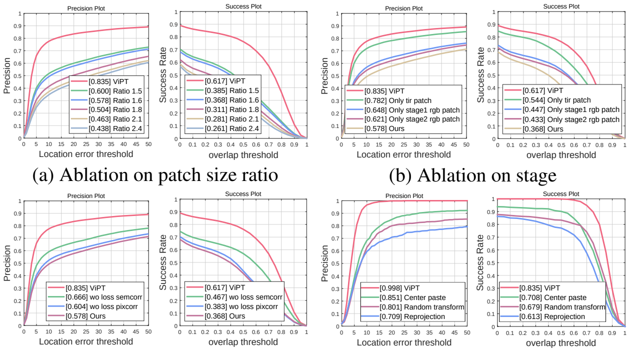


Figure 9: Practical application in the physical domain.



(a) Ablation on patch size ratio

(b) Ablation on stage

(b) Ablation on loss

(d) Ablation on reprojection

Figure 10: Quantitative comparison of ablation studies.

Conclusion

As the first attempt, we study cross-modal patch-based attack against RGB-T tracking. A coarse-to-fine strategy is proposed to progressively disturb RGB-T trackers. We then introduce pixel-level and semantic-level losses to interfere with feature responses. Finally, we present a point tracking-based reprojection strategy to enhance practical robust of patches in the real world. Experimental results demonstrate that our CMS can significantly degrade the performance of RGB-T trackers in both digital and physical domains.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62371350, U23B2050, and 62276192).

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision Workshops*, 850–865.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 6182–6191.

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 927–935.

Chen, X.; Yan, X.; Zheng, F.; Jiang, Y.; Xia, S.-T.; Zhao, Y.; and Ji, R. 2020. One-shot adversarial attacks on visual tracking with dual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10176–10185.

Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. Probabilistic regression for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7183–7192.

Ding, L.; Wang, Y.; Yuan, K.; Jiang, M.; Wang, P.; Huang, H.; and Wang, Z. J. 2021. Towards universal physical attacks on single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1236–1245.

Doersch, C.; Yang, Y.; Vecerik, M.; Gokay, D.; Gupta, A.; Aytar, Y.; Carreira, J.; and Zisserman, A. 2023. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE International Conference on Computer Vision*, 10061–10072.

Hu, Y.-C.-T.; Kung, B.-H.; Tan, D. S.; Chen, J.-C.; Hua, K.-L.; and Cheng, W.-H. 2021. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, 7848–7857.

Huang, X.; Miao, D.; Wang, H.; Wang, Y.; and Li, X. 2024. Context-Guided Black-Box Attack for Visual Tracking. *IEEE Transactions on Multimedia*, 26: 8824–8835.

Jia, S.; Song, Y.; Ma, C.; and Yang, X. 2021. 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*, 6709–6718.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2022a. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Li, P.; Chen, B.; Bai, L.; Qiao, L.; Li, B.; and Ouyang, W. 2022b. SiamSampler: Video-guided sampling for Siamese visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4): 1752–1761.

Li, Z.; Shi, Y.; Gao, J.; Wang, S.; Li, B.; Liang, P.; and Hu, W. 2021. A simple and strong baseline for universal targeted attacks on Siamese visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3880–3894.

Liu, L.; Li, C.; Xiao, Y.; Ruan, R.; and Fan, M. 2024. Rgbt tracking via challenge-based appearance disentanglement and interaction. *IEEE Transactions on Image Processing*, 33: 1753–1767.

Liu, L.; Li, C.; Xiao, Y.; and Tang, J. 2023. Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In *Proceedings of the ACM International Conference on Multimedia*, 3129–3137.

Long Li, C.; Lu, A.; Hua Zheng, A.; Tu, Z.; and Tang, J. 2019. Multi-adapter RGBT tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Lou, S.; Liu, B.; Bao, J.; Ding, J.; and Yu, J. 2023. Follow-me: Deceiving Trackers with Fabricated Paths. In *Proceedings of the ACM International Conference on Multimedia*, 8808–8818.

Nakka, K. K.; and Salzmann, M. 2022. Universal, transferable adversarial perturbations for visual object trackers. In *Proceedings of the European Conference on Computer Vision*, 413–429.

Nam, H.; and Han, B. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4293–4302.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2831–2838.

Yan, B.; Wang, D.; Lu, H.; and Yang, X. 2020. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 990–999.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27026–27035.

Zhang, F.; Peng, H.; Yu, L.; Zhao, Y.; and Chen, B. 2023. Dual-modality space-time memory network for RGBT tracking. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–12.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8886–8895.

Zhao, S.; Xu, T.; Wu, X.-J.; and Kittler, J. 2023. Pluggable Attack for Visual Object Tracking. *IEEE Transactions on Information Forensics and Security*, 19: 1227–1240.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9516–9526.