Semantic Feature Purification for Adversarially-Aware RGB-T Tracking

Jiahao Wang[1,2,3,4], **Fang Liu**[1,2,3,4*], **Hao Wang**[1,2,3,4], **Shuo Li**[1,2,3,4], **Xinyi Wang**[1,2,3,4], **Puhua Chen**[1,2,3,4]

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
[2]International Research Center for Intelligent Perception and Computation
[3]Joint International Research Laboratory of Intelligent Perception and Computation
[4]School of Artificial Intelligent, Xidian University, Xi'an,710071, P.R. China
jh_wang1024@163.com, f63liu@163.com

Abstract

RGB-T tracking is increasingly deployed in safety-critical applications such as autonomous driving, surveillance, and rescue robotics, where tracking reliability is essential under adverse conditions. Although the fusion of RGB and thermal infrared (TIR) modalities offers improved robustness in low-light and occluded scenes, recent findings show that RGB-T trackers remain highly susceptible to subtle input perturbations, human-imperceptible modifications that exploit cross-modal inconsistencies to mislead tracking outputs. In real-world scenarios, such perturbations can arise from sensor spoofing, infrared camouflage, or physical-world attacks, posing serious risks to operational safety. To address this, we propose SFPT, a Semantic Feature Purification framework that enhances RGB-T tracking at the representation level. Rather than filtering corrupted inputs at the pixel level, SFPT introduces task-specific semantic anchors into the feature space to reinforce perturbation-invariant cues. These anchors are derived from descriptive language, interact with visual features to purify representations. To further suppress modality-specific interference, we design an Adaptive Perturbation-Guided Cross-Modal Fusion (APG-CMF) module, which leverages language and visual signals to estimate reliability and dynamically reweight cross-modal features, ensuring robust fusion under perturbation conditions. Extensive experiments under diverse perturbation conditions validate the effectiveness of our approach. Notably, SFPT maintains performance comparable to clean settings even when subjected to perturbations of strength $\frac{1}{255}$ and $\frac{4}{255}$, demonstrating strong resilience to real-world interference.

Introduction

RGB-T tracking is increasingly used in safety-critical applications such as autonomous driving, surveillance, and military operations, where accurate and continuous target localization is vital (Xiao et al. 2022; Wang et al. 2024b; Hou et al. 2024; Tang et al. 2023b; Liu et al. 2025). By fusing RGB and thermal infrared (TIR) modalities, these systems remain effective in low visibility conditions like nighttime, fog, or occlusion. However, this cross-modal reliance also introduces security risks: recent studies show that RGB-T trackers are vulnerable to imperceptible adversarial perturbations that can severely mislead predictions (Mao et al. 2023; Schlarmann and Hein 2023). In practical adversarial scenarios, such manipulations can take diverse forms: interference can be introduced remotely by injecting signals into camera feeds (e.g., through wireless spoofing or infrared projection), physically deployed in the environment (e.g., using adversarial patches, reflective tape, or camouflage), or embedded digitally during data transmission and processing. These disturbances may selectively affect either RGB or TIR inputs, or target the misalignment between them, making the fusion mechanism itself a potential vulnerability. Even minimal perturbations can cause severe tracking degradation in real-time systems, such as autonomous vehicles failing to detect pedestrians or surveillance systems losing track of suspicious targets. Thus, defending RGB-T trackers against adversarial interference is not a theoretical luxury, it is a practical necessity for secure and trustworthy deployment in hostile or uncontrolled environments.

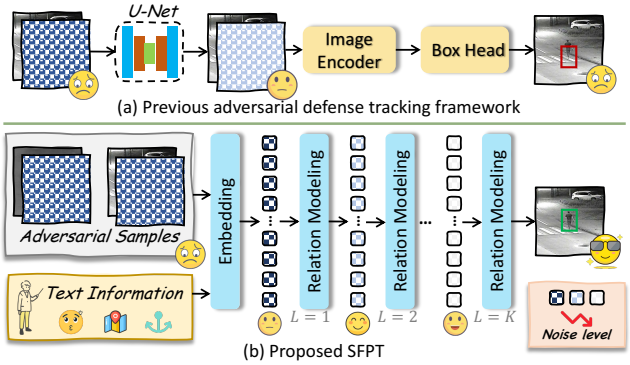

Figure 1: Comparison of different paradigm frameworks. (a) Previous adversarial defense tracking primarily focuses on RGB visual tracking, employing a U-Net architecture as the defense network. (b) The proposed SFPT incorporates text information into the defense network, exploring adversarial defense strategies tailored for RGB-T visual tracking.

Despite increasing interest in adversarial robustness for classification tasks, ensuring reliable RGB-T tracking under adversarial interference remains an open and technically demanding challenge. Existing defense methods fall broadly into three categories: (1) input-level filtering via models such as U-Net (Ronneberger, Fischer, and Brox 2015; Wu et al. 2025) (Figure 1(a)); (2) adversarial training using

*Corresponding author

perturbed inputs (Wei et al. 2023b); and (3) architectural adaptations designed to improve robustness. However, these techniques are predominantly designed for unimodal settings and exhibit critical shortcomings in multimodal scenarios like RGB-T tracking. First, input-level defenses assume spatially uniform distortions and overlook modality-specific interference, e.g., attacks targeting only the RGB stream while TIR remains clean, resulting in asymmetric degradation. Second, most approaches operate on individual frames and ignore temporal consistency, which is essential for detecting anomalies in motion and trajectory. Third, rigid fusion strategies (e.g., early or late fusion) fail to account for modality reliability, often amplifying the effect of corrupted inputs. Moreover, although recent vision-language models like MVTTG (Wang et al. 2024c) introduce text guidance for better representations, they focus on semantic enhancement rather than defense. To our knowledge, no existing work explicitly addresses multimodal feature purification for RGB-T tracking under adversarial threats.

We argue that to build truly reliable RGB-T trackers, adversarial defense must extend beyond input-level correction and instead operate in the feature space, where high-level semantics are structured and decision boundaries are formed. While adversarial perturbations are often subtle in the input domain, they can significantly distort the learned representations, especially by disrupting modality consistency or misaligning features with their true semantic meaning (Mai, Hu, and Xing 2020; Cui et al. 2024). However, language-derived descriptions provide a robust semantic prior that remains stable across input variations. By aligning visual features with these input-invariant semantic anchors, models can be guided back toward the correct semantic manifold, thereby improving resilience to adversarial manipulation. Furthermore, since adversarial interference often varies across modalities, fusion strategies should be adaptive and integrity-aware rather than statically coupled. These insights motivate a defense framework that purifies multimodal features via semantic alignment and dynamically reweights RGB-T contributions based on estimated input reliability.

In this work, we present SFPT, a semantic feature purification framework that enhances the adversarial robustness of RGB-T trackers through feature-space defense, as depicted in Figure 1(b). Unlike traditional approaches that attempt to suppress input-level distortions, SFPT adopts a modular, plug-in design that purifies deep features via semantic anchoring and adaptive cross-modal fusion. Specifically, we employ a pretrained vision-language model to generate textual prompts that describe the target object, and use these prompts to guide the purification of RGB and TIR features through a prompt-aware defense module. To further mitigate the impact of modality-specific perturbations, we introduce an Adaptive Perturbation-Guided Cross-Modal Fusion module that estimates the reliability of each modality and dynamically reweights their contributions during training. This architecture reduces reliance on compromised inputs and stabilizes the training dynamics under perturbations. In summary, our key contributions are as follows:

1. We propose SFPT, an adversarial defense framework that purifies semantic features for robust RGB-T tracking. Unlike conventional input-level defenses, SFPT aligns corrupted multimodal representations with language-guided semantic prompts, enabling effective feature-space purification against adversarial perturbations.
2. We introduce a text-driven, interference-aware purification strategy that injects semantic prompts into the tracking pipeline and adaptively reweights RGB and TIR features based on estimated input reliability. This unified mechanism allows SFPT to align features with text-conditioned semantics while dynamically suppressing corrupted signals from unreliable modalities.
3. We conduct extensive evaluations on RGB-T tracking benchmarks under various perturbation scenarios. Results show that SFPT consistently outperforms state-of-the-art trackers, maintaining high performance with minimal degradation even under severe input manipulation.

Related Work

RGB-T Tracking: Single-object tracking aims to locate a target in a video sequence continuously (Ye et al. 2022; Li et al. 2023b; Wei et al. 2023a; Chen et al. 2023; Wang et al. 2024a), with applications in surveillance (Henriques et al. 2014; Zhang et al. 2020), transportation (Ge et al. 2023), and autonomous driving (Fang et al. 2020; Hui et al. 2021). While RGB tracking has improved, its dependence on a single modality reduces reliability in challenging scenes. Fusing multiple modalities enhances robustness by leveraging complementary visual information (Shao et al. 2025).

The key challenge in RGB-T tracking is effectively merging different modalities. Early fusion methods, such as weighted averaging or selection mechanisms (Wu et al. 2011; Li et al. 2017; Jiao et al. 2023), are simple but do not fully exploit cross-modal complementarity. With deep learning advancements, neural network-based fusion techniques (Li et al. 2024; Wang et al. 2024b; Tang et al. 2025) now dominate, learning intrinsic cross-modal correlations for more precise and robust tracking. For instance, (Xiao et al. 2022) introduces an attribute-based progressive fusion network designed explicitly for RGB-T tracking, extracting and merging features through separate branches to effectively address five typical tracking challenges. In addition, (Zhu et al. 2023) takes another approach by adjusting pre-trained models using modality-related prompts, thus boosting performance across various downstream tracking tasks and showcasing new potential for multi-modal tracking. Another innovative model, (Hou et al. 2024) proposes a symmetric multi-modal tracking framework that combines lightweight adaptive strategies with self-distillation to improve tracking accuracy. Finally, (Wang et al. 2024c) enhances target and search area semantic information using language-based augmentation, demonstrating the potential of language descriptions in multi-modal visual tracking.

Adversarial Defense: In recent years, adversarial attacks have emerged as a critical threat to deep learning models, particularly in computer vision tasks such as classification, detection, and tracking (Athalye, Carlini, and Wagner 2018; Long et al. 2024; Baniecki and Biecek 2024; Li et al. 2025b,a; Zheng et al. 2025). To counter such threats, various

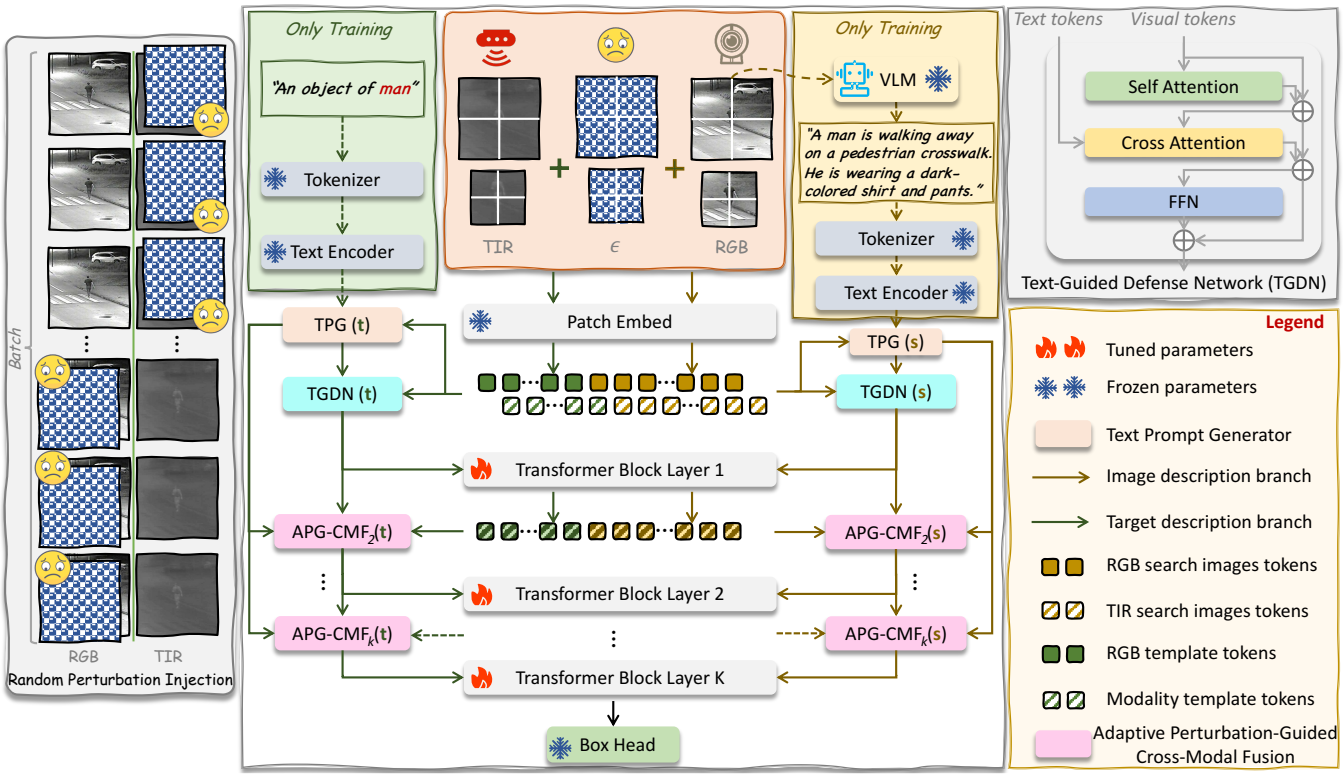

Figure 2: The overall structure of the proposed SFPT framework. This framework leverages text information to guide the defense network in mitigating adversarial perturbations that affect visual features. Here, "**s**" denotes the search image branch, and "**t**" represents the template image branch. The detailed architecture of TGDN is illustrated in the top-right corner.

defense strategies have been proposed, broadly categorized into three main paradigms: adversarial training, input transformation, and architectural modification. Adversarial training, which introduces adversarial examples during training to improve robustness, remains one of the widely adopted techniques (Madry et al. 2017; Wong, Rice, and Kolter 2020; Liang et al. 2022). However, most existing adversarial training methods are tailored to single-modality scenarios (e.g., RGB images), making them difficult to generalize effectively to multi-modal contexts such as RGB-T tracking, where modality-specific perturbations and cross-modal interactions must be jointly addressed (Liu et al. 2024). Input transformation methods suppress adversarial perturbations through image preprocessing techniques such as denoising, smoothing, or compression (Laykaviriyakul and Phaisangittisagul 2023; Tang et al. 2023a). While these methods are model-agnostic and straightforward, they often lack adaptivity and fail to preserve modality-aware semantics, thus limiting their effectiveness in multimodal settings. Model structure-based defenses improve robustness by modifying network components or incorporating specialized defense modules (Gosch et al. 2024; Wu et al. 2025). Though these approaches have shown promise in tasks like classification and RGB tracking, their extensions to RGB-T tracking remain limited. This is due to the added complexity of multi-modal feature fusion and the need to handle cross-modal adversarial inconsistencies. In summary, although significant advancements have been achieved in adversarial defense for single-modality tasks, effective solutions specifically designed for multi-modal tracking, especially in RGB-T scenarios, are still unavailable.

We propose an adversarial defensive strategy that improves the robustness of RGB-T tracking via visual-language interactions to tackle these issues. This approach leverages cross-modal characteristics to facilitate effective multi-modal information integration, diminishing dependence on any one modality. Furthermore, our adaptive perturbations management mechanism facilitates the dynamic segregation of hostile perturbations from tracking characteristics under assault, yielding a resilient defense efficacy.

Methodology

Problem Definition

Given a video sequence $\{\boldsymbol{I}_i\}_{i=1}^{n}$ with n consecutive frames, the goal of RGB-T tracking is to locate the target object in each subsequent frame based on the initial bounding box $\boldsymbol{B}_0$ and additional guidance, including text input. Since this is an RGB-T tracking task, the video frames consist of RGB frames $\{\boldsymbol{I}_i^{RGB}\}_{i=1}^{n}$ and thermal infrared (TIR) frames $\{\boldsymbol{I}_i^{TIR}\}_{i=1}^{n}$, with each modality contributing complementary information to enhance tracking robustness in complex environments. Our proposed framework processes these inputs through the **T**ext-**G**uided **D**efense **N**etwork (**TGDN**). TGDN aligns text and visual features while de-

fending against adversarial perturbations targeting RGB and TIR frames. The inputs to TGDN include the RGB frames $\{\boldsymbol{I}_i^{RGB}\}_{i=1}^n$, TIR frames $\{\boldsymbol{I}_i^{TIR}\}_{i=1}^n$, text information $\boldsymbol{T}$, and the initial bounding box $\boldsymbol{B}_0$.

The objective of the tracker is to learn a function f capable of predicting the target's bounding box $\boldsymbol{B}_i$ in each frame $\boldsymbol{I}_i$ while ensuring robustness against adversarial perturbations. The TGDN processes perturbed input frames $\hat{\boldsymbol{I}}_i^{RGB,atk}$ and $\hat{\boldsymbol{I}}_i^{TIR,atk}$, generating defended samples $\hat{\boldsymbol{I}}_i^{RGB}$ and $\hat{\boldsymbol{I}}_i^{TIR}$. The tracker uses these samples to produce robust tracking results. The adversarial defense objective is defined as:

$$\min_{\theta} L_{adv}\left(f(\boldsymbol{z}, Def(\hat{\boldsymbol{I}}_i^{RGB,atk}, \hat{\boldsymbol{I}}_i^{TIR,atk}, \theta)), \boldsymbol{y}_i^{reg}\right), \quad (1)$$

where $\boldsymbol{z}$ is the initial template, $\boldsymbol{y}_i^{reg}$ is the regression labels, and $Def(\cdot, \theta)$ represents the TGDN with parameters θ. The framework's output is the predicted bounding box $\boldsymbol{B}_i$ for the target object in each frame. TGDN integrates textual semantics into visual features, bolstering the model's robustness. Unlike traditional methods, TGDN leverages textual cues to refine decision-making in noisy environments, preventing perturbations' amplification. This approach enhances the synergy between adversarial defense and perturbation suppression, effectively empowering the model to manage adversarial perturbations and uncertainties in RGB-T tracking.

Overall

RGB-T tracking enhances the comprehensiveness of information by incorporating auxiliary visual modalities TIR. In this work, we extract image and category descriptions separately from the search and template images, ensuring that this text information is immune to adversarial perturbations. We introduce these textual descriptions into the defense network to improve decision-making. Moreover, we incorporate text information at each layer, enabling the model's context-aware capability to effectively distinguish the target from interference in complex or multi-object environments. This layer-wise approach also filters out perturbations inconsistent with the provided descriptions. As illustrated in Figure 2, the proposed SFPT framework consists of the following key components: Image Encoder, Text Encoder, TGDN, **T**ext **P**rompt **G**enerator (**TPG**) and **A**daptive **P**erturbation-**G**uided **C**ross-**M**odal **F**usion (**APG-CMF**). Here, the structure of the TPG is identical to that of the TGDN.

Given RGB and auxiliary visual images TIR, we first introduce input perturbations to the visual inputs. Specifically, for each batch of image pairs, perturbations are randomly added to $b\%$ of the RGB images, while the remaining $(100-b)\%$ of image pairs receive perturbations on their corresponding auxiliary visual images, as shown in Figure 2(left). This randomized perturbation injection strategy ensures a diverse and balanced distribution of interference across different modalities, enhancing the model's ability to handle a wide range of adversarial conditions. The image and text encoders project visual inputs and language descriptions, generated by a vision-language model (VLM) and category labels, into a shared feature space, producing semantically aligned image and text token embeddings. The TPG takes text features (text queries) and visual features as input to generate text prompts, which are then used to replace the original text tokens. Text prompts and visual features are input into TGDN to obtain the initial defense sample tokens $\boldsymbol{V}_{init} \in \mathbb{R}^{(N_z+N_s)\times C}$, where C represents embedding dimension, N_z and N_s represent the number of template and search tokens respectively. We denote the process as:

$$\boldsymbol{P}_T = f_{tp}(\boldsymbol{V}_{init}, \boldsymbol{F}_T), \quad (2)$$

$$\boldsymbol{V}_{init} = Def(\boldsymbol{V}_{init}, \boldsymbol{P}_T), \quad (3)$$

where $\boldsymbol{F}_T \in \mathbb{R}^{1\times C}$ and $\boldsymbol{P}_T \in \mathbb{R}^{1\times C}$ are the text features and text prompt tokens, $Def(\cdot,\cdot)$ and $f_{tp}(\cdot,\cdot)$ represent TGDN and TPG respectively. The initial defense sample tokens are then fed into Transformer Block layers, where text tokens are refined through the APG-CMF to further suppress cross-modal perturbation effects. The refined defense tokens are then added residually to the subsequent APG-CMF layer:

$$\boldsymbol{V}_{clean}^{i+1} = f_{AC}(\boldsymbol{V}^{i+1}, \boldsymbol{P}_T, \boldsymbol{V}_{clean}^{i}). \quad (4)$$

Here, $\boldsymbol{V}_{clean}^{i}$ represents the defense tokens at the i-th layer, and $f_{AC}(\cdot,\cdot,\cdot)$ is APG-CMF. Finally, the refined defense embeddings are passed through the Box Head to predict the target's bounding box precisely. Notably, during the training phase, the VLM, text encoder, and Box Head are kept frozen. In the image encoder, only the added Readapter (Luo et al. 2023) is optimized, while all other parameters remain fixed. This selective optimization strategy ensures that the core components retain their pre-trained capabilities, while the Readapter learns to adapt the visual features more effectively for the tracking task. The alignment between the generated text prompts and the original text features can be formulated as a cosine similarity loss to ensure consistency. The cosine similarity loss $\mathcal{L}_{\text{cos}}$ is defined as:

$$\mathcal{L}_{\text{cos}} = 1 - \frac{\langle \mathbf{P_T}, \mathbf{F_T} \rangle}{\|\mathbf{P_T}\|_2 \cdot \|\mathbf{F_T}\|_2}, \quad (5)$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product between the two feature vectors. $\|\cdot\|_2$ represents the L_2-norm. The loss $\mathcal{L}_{\text{cos}}$ ensures that the generated prompts maintain high semantic similarity with the original text tokens by minimizing their angular distance in the feature space.

Adaptive Perturbation-Guided Cross-Modal Fusion

APG-CMF addresses the challenge of maintaining tracking performance under cross-modal interference by leveraging visual and textual information, as shown in Figure 3. The method enhances multimodal fusion by estimating perturbation intensity and dynamically weighting features to mitigate modality-specific disruption. Given text prompt tokens $\boldsymbol{P}_T$ and visual tokens $\boldsymbol{V}_{clean}^{i}$ and $\boldsymbol{V}^{i+1}$ from two stages, the module first divides each set of visual features along the channel dimension into two parts. The split features are then transformed via individual 1×1 convolutions, summed, and fused. A subsequent 1×1 convolution restores the original channel dimensionality, producing the fused representation $\boldsymbol{V}_{fuse}^{i+1}$. This design efficiently integrates subspace information, enhancing the diversity and reliability of the combined visual representation. Next, the fused visual tokens are reshaped to $N_V \times C$ and passed through a 3×3 convolution, followed by ReLU activation and average pooling, to extract a compact representation of modality-specific perturbation. This representation is then projected via a linear layer to obtain an interference estimation vector $\boldsymbol{E}_{perturb}$. Based on the

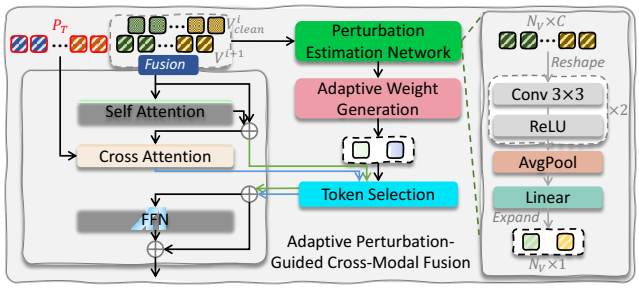

Figure 3: The detailed structure of the proposed APG-CMF. It leverages the interaction between text prompts and visual features, combined with a perturbation estimation network and an adaptive weight generation module, to achieve dynamic feature weighting and effective suppression of modality-specific interference.

estimated perturbation, the attention weights between visual and language tokens are adaptively modulated, enabling responsive cross-modal fusion as follows:

$$\boldsymbol{W}_{perturb} = \frac{1}{1 + e^{-\beta \cdot (\boldsymbol{E}_{perturb} - \gamma)}}, \tag{6}$$

where $\boldsymbol{W}_{perturb}$ (it is represented by $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$) is perturbation-aware adaptive weights, and β and γ are hyperparameters set to 1 and 0.5, respectively.

Self-Attention captures intra-modal dependencies, while Cross-Attention aligns visual and textual features, ensuring coherent feature interaction. Token selection based on perturbation-aware weights ensures that only relevant tokens contribute to the final fused representation. The fusion is performed using a weighted combination:

$$\boldsymbol{V}_{ts}^{i+1} = \boldsymbol{w}_0 \cdot SA(\boldsymbol{V}_{fuse}^{i+1}) + \boldsymbol{w}_1 \cdot CA(SA(\boldsymbol{V}_{fuse}^{i+1}), \boldsymbol{P}_T), \tag{7}$$

$$\boldsymbol{V}_{clean}^{i+1} = FFN(\boldsymbol{V}_{ts}^{i+1}) + \boldsymbol{V}_{ts}^{i+1}. \tag{8}$$

Here, $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ represent the weights of token selection. The fused tokens are then processed by a Transformer Block with residual connections to refine the features across layers. The core of the perturbation-aware fusion mechanism lies in leveraging interference estimation to adaptively modulate the attention mechanism: *when perturbation intensity is high, the model places greater emphasis on language-derived semantics; conversely, under cleaner conditions, it prioritizes visual information for tracking decisions.*

Optimization

During the optimization process of SFPT, adversarial loss is employed to enhance the model's robustness. In training, the goal is to maximize the internal loss to identify adversarial samples (inputs that cause the greatest deviation in the model's predictions) and minimize the external objective by updating the model parameters, ensuring robustness against these adversarial samples. Specifically, each training batch involves two forward passes: the first pass computes the adversarial loss $L_{adv}^{(1)}$ using the initial perturbation, followed by an update to generate the adversarial perturbation. In the second forward pass, the new adversarial loss $L_{adv}^{(2)}$ is calculated. Finally, the ADAM optimizer (Kingma 2014) minimizes $L_{adv}^{(2)}$ to update the network parameters, ensuring the model can reliably track targets even in noisy environments. The overall process is shown in Algorithm 1.

Algorithm 1: Framework of the adversarial training process of proposed SFPT on search branch

Require: training dataset D, training epochs N, batch size B, perturbation budget ε

Ensure: trained text-guided defense network parameters θ

for int i=1 to N **do**
  for sample $\{\boldsymbol{z}, \boldsymbol{I}^{RGB}, \boldsymbol{I}^{M}, \boldsymbol{T}, \boldsymbol{y}_{reg}\} \in D$ **do**
    Initialize training perturbation δ with the gaussian noise. $\delta \sim U[-\epsilon, \epsilon]$
    Obtain the defense sample $\hat{\boldsymbol{I}}^{RGB}$ and $\hat{\boldsymbol{I}}^{TIR}$ by feeding $\boldsymbol{I}^{RGB} + \delta$ and $\boldsymbol{I}^{TIR} + \delta$ into the text-guided defense network parameterized by θ.
    $\hat{\boldsymbol{I}}^{RGB}, \hat{\boldsymbol{I}}^{TIR} = Def(\boldsymbol{I}^{RGB} + \delta, \boldsymbol{I}^{TIR} + \delta, \theta)$
    After the first forward pass, calculate the Adv-Loss with $\hat{\boldsymbol{I}}^{RGB}$ and $\hat{\boldsymbol{I}}^{TIR}$.
    $L_{adv}(f(\boldsymbol{z}, \hat{\boldsymbol{I}}^{RGB}, \hat{\boldsymbol{I}}^{TIR}), \boldsymbol{y}_{reg})$
    Update δ to adversarial perturbation δ^{adv}.
    $\theta_{adv} = \theta + \epsilon \cdot sign(\bigtriangledown_{\{\boldsymbol{I}^{RGB}, \boldsymbol{I}^{TIR}\}} L_{adv}))$
    Reobtain the defense sample $\hat{\boldsymbol{I}}^{RGB}$ and $\hat{\boldsymbol{I}}^{TIR}$ by feeding the search region with adversarial perturbation $\boldsymbol{I}^{RGB} + \delta^{adv}$ and $\boldsymbol{I}^{TIR} + \delta^{adv}$ into the defense network parameterized by θ.
    $\hat{\boldsymbol{I}}^{RGB}, \hat{\boldsymbol{I}}^{TIR} = Def(\boldsymbol{I}^{RGB} + \delta, \boldsymbol{I}^{TIR} + \delta, \theta)$
    After the second forward pass, recalculate the L_{adv} with new $\hat{\boldsymbol{I}}^{RGB}$ and $\hat{\boldsymbol{I}}^{TIR}$.
    $L_{adv}(f(\boldsymbol{z}, \hat{\boldsymbol{I}}^{RGB}, \hat{\boldsymbol{I}}^{TIR}), \boldsymbol{y}_{reg})$
    Compute the gradient of L_{adv} to defense network parameters θ and update θ with ADAM optimizer.
  end for
end for

The overall loss function of SFPT integrates multiple components to ensure accurate tracking and robustness in noisy environments. The overall loss L_{total} is:

$$L_{total} = \lambda_1 L_1 + \lambda_{iou} L_{iou} + \lambda_{cos} L_{cos}, \tag{9}$$

where L_1 represents the bounding box regression loss, L_{iou} denotes the generalized IoU loss, and L_{cos} corresponds to the semantic consistency loss for text prompts. The weights λ_1, λ_{iou}, and λ_{cos} are set to 5, 2, and 1, respectively, to balance the contributions of each loss component.

Experiments

Implementation Details

Model: The proposed SFPT framework is constructed with a VIT-B (Dosovitskiy et al. 2020) encoder of 12 Transformer Block Layers. We place three APG-CMF modules at the encoder's 2nd, 6th, and 12th input layers. For VLM, we use BLIP2 (Li et al. 2023a) to generate image descriptions, while the text encoder is based on RoBERTa-Base (Liu 2019). The tracker's speed remains similar to the original foundation model (Wu et al. 2023).

Training Details: Following an end-to-end training and testing approach, like other RGB-T trackers, SFPT is trained on

Dataset		FFT	$\epsilon = \frac{1}{255}$						$\epsilon = \frac{4}{255}$					
		FM	FM			**Ours**			FM			**Ours**		
		-	Search	Template	Both	Search	Template	Both	Search	Template	Both	Search	Template	Both
LasHeR	PR(↑)	69.0	64.0	63.5	65.2	69.4	69.6	70.4	58.8	58.6	57.6	70.2	68.5	68.4
(Li et al. 2021)	SR(↑)	55.1	49.8	49.2	50.8	55.6	55.8	56.5	45.0	44.9	44.1	56.2	55.0	55.0
$\triangle_{\mathbf{gap}}$		-	-5.15	-5.7	-4.05	+0.45	+0.65	+1.4	-10.85	-10.3	-11.15	+1.15	-0.3	-0.35
RGBT234	MPR(↑)	81.5	77.9	78.5	78.3	84.9	83.9	85.5	75.9	75.5	75.9	85.8	85.5	85.2
(Li et al. 2019)	MSR(↑)	59.2	55.7	56.0	56.2	63.0	62.4	63.4	53.8	54.1	54.1	63.7	62.6	62.4
$\triangle_{\mathbf{gap}}$		-	-3.55	-3.1	-3.1	+3.6	+2.8	+4.1	-5.5	-5.55	-5.35	+4.4	+3.7	+3.45
RGBT210	PR(↑)	78.4	75.8	76.5	76.2	83.0	82.8	83.4	74.7	74.4	74.9	82.6	81.9	82.5
(Li et al. 2017)	SR(↑)	54.4	52.9	53.3	53.3	59.5	59.1	60.0	51.3	51.1	51.5	58.2	57.7	57.8
$\triangle_{\mathbf{gap}}$		-	-2.05	-1.5	-1.65	+4.85	+4.55	+5.3	-3.4	-3.65	-3.2	+4	+3.65	+3.75
GTOT	PR(↑)	91.1	86.2	85.5	85.9	92.2	92.0	93.3	83.9	82.8	83.1	91.7	91.5	92.1
(Li et al. 2016)	SR(↑)	71.9	65.1	64.9	65.0	74.8	74.4	75.9	62.7	61.8	62.4	75.0	74.5	75.1
$\triangle_{\mathbf{gap}}$		-	-5.85	-6.3	-6.05	+2	+1.7	+3.1	-8.2	-9.2	-8.75	+1.85	+1.75	+2.1

Table 1: Performance of the foundation model (FM) and the proposed SFPT across different levels of perturbation ϵ, evaluated on multiple RGB-T tracking datasets. "FFT" represents the full fine-tuned results of the foundation model on clean samples, while $\triangle_{\mathbf{gap}}$ indicates the average change across two metrics between the defense performance of the proposed SFPT and FM at various values of ϵ compared to the original results. Improvements are highlighted in **red**, while declines are shown in **blue**.

LasHeR (Li et al. 2021) for RGB-T tracking. The model is trained on two NVIDIA 4090 GPUs. We use DropTrack (Wu et al. 2023) as the foundation tracker during training, employing the AdamW optimizer for 60 epochs with a weight decay of 10^{-4} and a batch size of 32. The initial learning rate is set to 4×10^{-4}, reducing by a factor of 10 after 48 epochs. The template image size is 192×192, and the search image size is 384×384. We limit the maximum number of text tokens for language input to 77. Adversarial training is conducted with perturbation budgets of $\epsilon = \frac{1}{255}$ and $\frac{4}{255}$. For the defense networks in both the template and search branches, we use an identical architecture, allowing for simultaneous deployment across branches or selective activation within a single branch. Additionally, to ensure the fairness of the experiments, the foundation model is full fine-tuned on the LasHeR dataset.

Comparative Experiments

Robustness Analysis. To assess the robustness of SFPT, we applied adversarial interference attacks separately to the template image and the search image and then to both images simultaneously to explore the tracker's resilience. Results in Table 1 demonstrate that SFPT consistently outperforms the foundation model (FM) across all evaluated perturbation levels and attack strategies. The significant performance gains ($\triangle_{\mathbf{gap}}$) achieved by SFPT over the FM, especially in high-interference scenarios, underscore the effectiveness of our defense strategy. This enhanced performance is attributed to SFPT's adaptive defense mechanism, which dynamically mitigates the impact of adversarial perturbation on critical tracking features, reinforcing its stability and resilience across diverse RGB-T datasets.

Performance comparison under different adversarial perturbations. Table 2 highlights SFPT's robustness against various adversarial attacks in RGB-T tracking, consistently outperforming the foundation model (FM). Notably, SFPT shows strong resilience to both natural noise and adversarial perturbations, with the highest gains under Exponential and Rayleigh noise, demonstrating its ability to suppress interference while preserving tracking accuracy. The consistently superior performance, especially on GTOT, underscores its adaptability to real-world scenarios. These results confirm that SFPT enhances feature discrimination and reinforces multimodal robustness, making it a reliable solution for adversarially robust tracking.

Exploration Studies

In this section, we examine the various components of SFPT. Unless stated otherwise, search and template images are perturbed by $\epsilon = \frac{1}{255}$ by default. For all ablations, we report performance using the RGB-T tracking benchmark LasHeR.

Adversarial Sample Generation Strategy and the Impact of ANF-CMF. In Table 3(#2 and #3), we evaluate the impact of two core components in SFPT: the randomized perturbation injection strategy and the APG-CMF module. As discussed in Methodology Section 3.2, the perturbation injection strategy diversifies and balances adversarial disturbance across different modalities. This helps prevent any single modality from being disproportionately affected by targeted perturbations, thereby improving the model's robustness in varied attack scenarios. This effect is evident in Row #2 of Table 3, where removing this strategy and applying uniform perturbation to all inputs leads to a noticeable performance drop, highlighting the importance of perturbation diversity for stable multi-modal tracking. Additionally, the APG-CMF module, which strengthens cross-modal fusion by estimating perturbation severity and dynamically reweighting modal contributions, also proves essential. Its removal (#3) results in further performance degradation, confirming APG-CMF's role in mitigating adversarial influence and enhancing semantic integration.

Impact of Text Information. Text information plays a critical role in SFPT's performance, as confirmed by the ablation studies shown in Table 3(#4 and #5). To analyze this impact, we first replaced the text information with learnable parameters; as observed in Figure 4(b)(Corresponds to #4), re-

Attack Method	LasHeR				RGBT234				GTOT			
	SR(↑)		PR(↑)		MSR(↑)		MPR(↑)		SR(↑)		PR(↑)	
	FM	SFPT (Ours)	FM	SFPT (Ours)	FM	SFPT (Ours)	FM	SFPT (Ours)	FM	SFPT (Ours)	FM	SFPT (Ours)
Gaussian noise	44.1	55.0 (+10.9)	57.6	68.4 (+10.8)	54.1	62.4 (+8.3)	75.9	85.2 (+9.3)	62.4	75.1 (+12.7)	83.1	92.1 (+9.0)
Uniform noise	54.1	55.1 (+1.0)	67.8	68.5 (+0.7)	58.0	63.8 (+5.8)	80.7	85.6 (+4.9)	60.9	70.8 (+9.9)	79.3	89.5 (+10.2)
Quantitative noise	54.6	55.6 (+1.0)	68.9	69.5 (+0.6)	59.2	64.8 (+5.6)	81.7	86.4 (+4.7)	61.8	71.0 (+9.2)	80.7	89.5 (+8.8)
Rayleigh Noise	43.0	50.6 (+7.6)	53.6	63.6 (+10.0)	45.6	55.1 (+9.5)	63.8	75.2 (+11.4)	59.2	65.7 (+6.5)	75.6	87.7 (+12.1)
Exponential noise	45.2	52.4 (+7.2)	56.6	66.3 (+9.7)	47.8	56.3 (+8.5)	66.7	78.2 (+11.5)	60.4	69.6 (+9.2)	76.1	82.6 (+6.5)
FGSM (Goodfellow et al. 2014)	52.1	54.9 (+2.8)	67.2	68.6 (+1.4)	59.3	64.0 (+4.7)	79.9	84.8 (+4.9)	60.6	65.8 (+5.2)	73.7	85.8 (+12.1)
PGD (Madry 2017)	21.3	30.2 (+8.9)	31.6	38.3 (+6.7)	35.0	43.6 (+8.6)	54.7	63.0 (+8.3)	32.1	40.0 (+7.9)	39.6	58.9 (+19.3)
CSA (Yan et al. 2020)	54.0	55.7 (+1.7)	68.1	69.3 (+1.2)	56.5	63.0 (+6.5)	80.3	84.9 (+4.6)	61.7	73.1 (+11.4)	78.8	88.8 (+10.0)
IoU Attack (Jia et al. 2021)	29.5	43.3 (+13.8)	36.4	55.0 (+18.6)	43.1	51.9 (+8.8)	61.4	70.2 (+8.8)	40.9	55.8 (+14.9)	59.7	75.7 (+16.0)
no Attack	**55.1**		**69.0**		**59.2**		**81.5**		**71.9**		**91.1**	

Table 2: The performance of SFPT on various RGB-T tracking datasets under different input perturbations, all with a perturbation intensity of $\epsilon = \frac{4}{255}$. "no Attack" represents the full fine-tuned (FFT) performance of the foundation model (FM). **Green** indicates the change in defense performance of SFPT relative to the attack FM.

#	Method	PR(↑)	SR(↑)
1	SFPT	**70.4**	**56.5**
2	w/o random perturbation injection	$70.1_{-0.3}$	$56.2_{-0.3}$
3	w/o APG-CMF	$66.7_{-3.7}$	$53.3_{-3.2}$
4	w/o Text information	$66.2_{-4.2}$	$53.5_{-3.0}$
5	only U-Net defense network	$62.9_{-7.5}$	$50.1_{-6.4}$
6	w/o TPG	$69.1_{-1.3}$	$55.5_{-1.0}$
7	$w/$ MSE loss	$69.9_{-0.5}$	$56.1_{-0.4}$
8	$w/$ CE loss	$69.5_{-0.9}$	$55.4_{-1.1}$

Table 3: Ablation Study of the Proposed SFPT on Different Components in the LasHeR dataset.

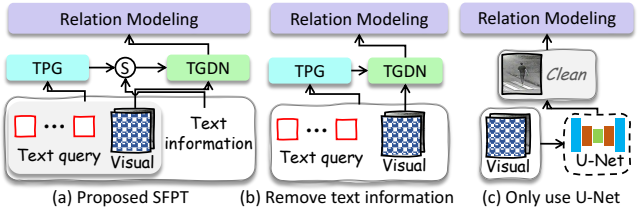

Figure 4: Diagram of the proposed SFPT and its variants. "S" represents similarity. (a) shows a schematic of SFPT, (b) removes text information from (a) and replaces it with a learnable text query without supervision from the original text, and (c) uses only the U-Net for perturbation removal.

moving text guidance leads to a decline in performance, underscoring its importance for robust tracking. Furthermore, we removed text information entirely and substituted the defense network with U-Net, a common choice in existing approaches (Wu et al. 2025), as shown in Figure 4(c). Row #5 demonstrates that this change results in a significant drop in performance, highlighting both text information's indispensable role and standard architectures' limitations in achieving effective multi-modal defense.

Different Semantic Consistency Losses. The ablation study results in Table 3 illustrate the TPG's effectiveness and the loss function's choice in SFPT's performance. As discussed in Methodology Section 3.2, TPG-generated text prompts are used during training instead of original text information, reducing the computational load of the tracking process by eliminating the need for continuous encoding by the text encoder. As shown in Row #6, the removal of TPG leads to a slight performance decrease, indicating that the generated prompts reduce computational requirements and are better adapted to task-specific data than the original text features. Additionally, we examine the impact of substituting the cosine similarity loss with alternative loss functions. In Rows #7 and #8, the cosine similarity loss is replaced with MSE and CE losses, resulting in noticeable performance declines. This outcome highlights the suitability of cosine similarity for aligning the multi-modal features in SFPT.

Conclusion

This work presents SFPT, a feature-level adversarial defense framework tailored for RGB-T tracking. By integrating a Text Prompt Generator (TPG) and an Adaptive Perturbation-Guided Cross-Modal Fusion (APG-CMF) module, SFPT leverages language-guided prompts and cross-modal consistency to suppress adversarial disruptions. This design enables dynamic disentanglement of modality-specific interference while maintaining compatibility with a wide range of transformer-based trackers without full retraining. Extensive experiments on standard RGB-T benchmarks show that SFPT consistently preserves performance under diverse perturbation conditions, highlighting its effectiveness and transferability for real-world, security-critical applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China(No.62576264), Project supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (No.2025ZD0551500, No.2025ZD0551502), the Key Project of National Natural Science Foundation of China (62431020,62231027), the Joint Fund Project of National Natural Science Foundation of China (No.U22B2054), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No.B07048), the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) (No.GZC20232033), the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT 15R53), the Key Scientific Technological Innovation Research Project by Ministry of Education and the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No.HMHAI-202404, No. HMHAI-202405).

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proc. of ICML*.

Attias, I.; Kontorovich, A.; and Mansour, Y. 2022. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*.

Baniecki, H.; and Biecek, P. 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proc. of CVPR*.

Cui, X.; Aparcedo, A.; Jang, Y. K.; and Lim, S.-N. 2024. On the robustness of large multimodal models against image adversarial attacks. In *Proc. of CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Z.; Zhou, S.; Cui, Y.; and Scherer, S. 2020. 3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud. *IEEE Sensors Journal*.

Ge, D.-y.; Yao, X.-f.; Xiang, W.-j.; and Chen, Y.-p. 2023. Vehicle detection and tracking based on video image processing in intelligent transportation system. *Neural Computing and Applications*.

Goodfellow et al. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gosch, L.; Geisler, S.; Sturm, D.; Charpentier, B.; Zügner, D.; and Günnemann, S. 2024. Adversarial training for graph neural networks: Pitfalls, solutions, and new directions. *Proc. of NeurIPS*.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*.

Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. 2024. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proc. of CVPR*.

Hui, L.; Wang, L.; Cheng, M.; Xie, J.; and Yang, J. 2021. 3d siamese voxel-to-bev tracker for sparse point clouds. *Proc. of NeurIPS*.

Jia, S.; Song, Y.; Ma, C.; and Yang, X. 2021. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *Proc. of CVPR*.

Jiao, L.; Zhang, X.; Liu, X.; Liu, F.; Yang, S.; Ma, W.; Li, L.; Chen, P.; Feng, Z.; Guo, Y.; et al. 2023. Transformer meets remote sensing video detection and tracking: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laykaviriyakul, P.; and Phaisangittisagul, E. 2023. Collaborative Defense-GAN for protecting adversarial attacks on classification system. *Expert Systems with Applications*.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*.

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

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proc. of ACM MM*.

Li, F.; Wang, T.; Zhu, L.; Li, J.; and Shen, H. T. 2025a. Attack as Defense: Proactive Adversarial Multi-modal Learning to Evade Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of ICML*.

Li, X.; Jiao, L.; Zhu, H.; Huang, Z.; Liu, F.; Li, L.; Chen, P.; and Yang, S. 2023b. A complex-former tracker with dynamic polar spatio-temporal encoding. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *Proc. of ECCV*.

Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Liang, H.; He, E.; Zhao, Y.; Jia, Z.; and Li, H. 2022. Adversarial attack and defense: A survey. *Electronics*.

Liu, F.; Wang, J.; Jiao, L.; Zhang, J.; Wang, H.; Li, S.; Li, L.; Chen, P.; Liu, X.; Ma, W.; et al. 2025. Remote Sensing Video Tracking: Current Status, Challenges and Future. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Long, J.; Jiang, T.; Yao, W.; Jia, S.; Zhang, W.; Zhou, W.; Ma, C.; and Chen, X. 2024. PapMOT: Exploring Adversarial Patch Attack Against Multiple Object Tracking. In *Proc. of ECCV*.

Luo, G.; Huang, M.; Zhou, Y.; Sun, X.; Jiang, G.; Wang, Z.; and Ji, R. 2023. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*.

Madry, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *stat*.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proc. of AAAI*.

Mao, C.; Geng, S.; Yang, J.; Wang, X.; and Vondrick, C. 2023. Understanding Zero-shot Adversarial Robustness for Large-Scale Models. In *Proc. of ICLR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*.

Schlarmann, C.; and Hein, M. 2023. On the adversarial robustness of multi-modal foundation models. In *Proc. of ICCV*.

Shao, Z.; Hu, Y.; Fan, B.; and Liu, H. 2025. PURA: Parameter Update-Recovery Test-Time Adaption for RGB-T Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Tang, X.; Yin, P.; Zhou, Z.; and Huang, D. 2023a. Adversarial perturbation elimination with GAN based defense in continuous-variable quantum key distribution systems. *Electronics*.

Tang, Z.; Xu, T.; Li, H.; Wu, X.-J.; Zhu, X.; and Kittler, J. 2023b. Exploring fusion strategies for accurate RGBT visual object tracking. *Information Fusion*.

Tang, Z.; Xu, T.; Wu, X.-J.; Zhu, X.; Cheng, C.; Feng, Z.; and Kittler, J. 2025. Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution. *IEEE Transactions on Image Processing*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*.

Wang, J.; Liu, F.; Jiao, L.; Gao, Y.; Wang, H.; Li, L.; Chen, P.; Liu, X.; and Li, S. 2024a. Satellite Video Object Tracking based on Location Prompts. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, J.; Liu, F.; Jiao, L.; Gao, Y.; Wang, H.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024b. Visual and Language Collaborative Learning for RGBT Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, J.; Liu, F.; Jiao, L.; Wang, H.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024c. Multi-modal visual tracking based on textual generation. *Information Fusion*.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023a. Autoregressive visual tracking. In *Proc. of CVPR*.

Wei, X.; Huang, Y.; Sun, Y.; and Yu, J. 2023b. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*.

Wu, Q.; Yang, T.; Liu, Z.; Wu, B.; Shan, Y.; and Chan, A. B. 2023. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proc. of CVPR*.

Wu, Y.; Blasch, E.; Chen, G.; Bai, L.; and Ling, H. 2011. Multiple source data fusion via sparse representation for robust visual tracking. In *14th International Conference on Information Fusion*.

Wu, Z.; Yu, R.; Liu, Q.; Cheng, S.; Qiu, S.; and Zhou, S. 2025. Enhancing Tracking Robustness with Auxiliary Adversarial Defense Networks. In *Proc. of ECCV*.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based progressive fusion network for rgbt tracking. In *Proc. of AAAI*.

Yan, B.; Wang, D.; Lu, H.; and Yang, X. 2020. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. In *Proc. of CVPR*.

Yan, S.; Yang, J.; Käpylä, J.; Zheng, F.; Leonardis, A.; and Kämäräinen, J.-K. 2021. Depthtrack: Unveiling the power of rgbd tracking. In *Proc. of ICCV*.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proc. of ECCV*.

Zhang, N.; Wu, C.; Wu, Y.; and Xiong, N. N. 2020. An improved target tracking algorithm and its application in intelligent video surveillance system. *Multimedia Tools and Applications*.

Zheng, Y.; Zhong, B.; Liang, Q.; Zhang, S.; Li, G.; Li, X.; and Ji, R. 2025. Towards universal modal tracking with online dense temporal token learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *Proc. of CVPR*.