OCTA-SOT: Online Cross-Modal Trajectory Adjustment for RGBT Anti-UAV Single Object Tracking under Spatio-Temporal Misalignment

Xiaokang Liu[1], Qi Jia[1] ⋆, Jinrui Wang[1], Chengzhou Li[1], Yu Liu[1], and Weimin Wang[1]

Dalian University of Technology, Dalian, China

Abstract. Due to the extensive application of UAVs in modern warfare, anti-UAV single object tracking has garnered increasing research interest. However, prevailing multi-modal methods heavily rely on the assumption of perfectly calibrated sensors, leaving the critical challenge of tracking under realistic spatio-temporal misalignment largely unaddressed. To tackle this issue, we formally introduce the novel task of uncalibrated RGBT anti-UAV tracking and mathematically formulate its observation process as a unified state-space model. Based on this theoretical foundation, we propose OCTA-SOT, *an* ***O****nline* ***C****ross-modal* ***T****rajectory* ***A****djustment* ***S****ingle* ***O****bject* ***T****racking framework*. As a completely training-free and plug-and-play module, it can be directly applied to off-the-shelf base trackers. During inference, it adaptively selects reliable trajectory information to update inter-modality mappings in real time, while simultaneously calibrating unconfident trackers and their results. At its core, a customized Kalman-driven mechanism dynamically adjusts these mappings, thereby achieving robust target tracking. Extensive experiments on the challenging Anti-UAV300 dataset demonstrate the exceptional effectiveness of our method. Without requiring any training, OCTA-SOT effectively enhances the DIMP tracker, achieving absolute improvements of 9.3% in AUC, 12.9% in Precision, and 11.9% in Normalized Precision, while simultaneously outperforming a series of current state-of-the-art methods. The source code is publicly available at https://github.com/xkliu-eps/AntiUAV_RGBT_Tracking/.

Keywords: Anti-UAV single object tracking · spatio-temporal misalignment · online joint correction · RGBT tracking

1 Introduction

The prominence of UAVs in modern battlefields has made anti-UAV systems a critical security priority [14, 40]. Central to these systems is real-time object tracking, which enables trajectory analysis and motion prediction [4, 19]. To handle complex environments, multi-modal fusion has gained traction for its complementary benefits [5, 32]. However, RGBT-based anti-UAV tracking remains relatively under-explored compared to other modalities.

⋆ Corresponding author: `jiaqi@dlut.edu.cn`

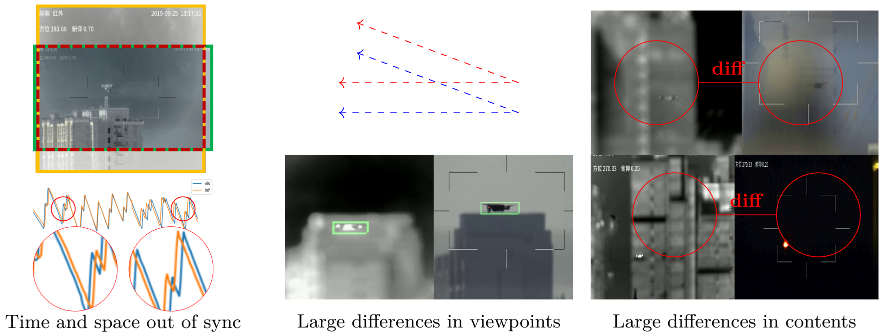

Fig. 1: Challenges in real-time RGBT anti-UAV tracking.

Real-time RGBT UAV tracking faces critical modal misalignment challenges. As illustrated in Fig. 1, heterogeneous sensors often exhibit inconsistent resolutions and unaligned scales (spatially), and asynchronous frame capture (temporally). Geometrically, inherent parallax causes target positions to be inconsistent across modalities; for instance, a UAV might appear above a rooftop in the visible frame but below it in the infrared (Fig. 1, middle). Furthermore, the sensors capture **inconsistent content**: visible images may contain cluttered backgrounds invisible to infrared [5, 13], whereas visible sensors fail in adverse lighting while infrared provides stable observations (Fig. 1, right) [12, 14]. These multifaceted factors render conventional "align-then-fuse" strategies [7, 34] ineffective. Such methods rely on static assumptions—such as fixed homography matrices for spatial registration or strictly synchronized hardware frame rates—which inevitably break down in dynamic UAV scenarios involving fast-moving targets, varying depths, and sensor jitter. Consequently, developing a framework to adaptively correct these misalignments without offline pre-calibration is paramount for reliable RGBT anti-UAV systems [4, 43].

Despite the urgent demand for robust tracking frameworks capable of operating under practical spatio-temporal misalignments, current RGBT single object tracking methodologies largely remain confined to the idealized premise of perfectly synchronized and spatially aligned modalities. Guided by this strict assumption, a variety of advanced RGBT trackers have been developed in recent years. For instance, in the realm of Siamese architectures, SiamCDA [37] reduces modal differences through complementarity-aware fusion, while MANet [17] leverages multiple adapters for hierarchical feature representation. More recently, Transformer-based frameworks such as TBSI [13] and BAT [5] have been proposed to enable deep, global cross-modal interactions. While these state-of-the-art methods achieve remarkable performance on highly aligned general benchmarks (e.g., LasHeR and RGBT234), their rigid alignment assumption completely disregards the widespread spatio-temporal and geometric deviations inherent in real-world anti-UAV systems, such as sensor latency, asynchronous frame rates, and uncalibrated hardware offsets. Consequently, directly applying these existing general RGBT trackers to practical, uncalibrated anti-UAV tasks remains exceedingly difficult.

To break through this bottleneck, we establish a formal definition for tracking under spatio-temporal misalignment and provide its concrete implementation. Formally, we first define the general paradigms of spatio-temporal misalignment and the standard tracking process: the former is formulated as a mapping process containing time-varying factors, while the latter is modeled as a state transition process of the tracker's internal parameters. By integrating these two aspects, we derive a unified state transition process that encompasses both the cross-modal mapping and the tracker parameters, driven by external video frame inputs. To instantiate this theoretical formulation, we propose OCTA-SOT (***O**nline **C**ross-modal **T**rajectory **A**djustment for RGBT anti-UAV **S**ingle **O**bject **T**racking*). Designed as a lightweight and training-free plug-and-play module, OCTA-SOT seamlessly mitigates inter-sensor deviations without requiring any retraining of the base trackers. Specifically, the framework continuously maintains historical tracking outputs to evaluate the real-time reliability of each modality. Guided by these data-driven decisions, it performs adaptive cross-modal state correction—elegantly projecting high-confidence tracking states to rescue degraded modalities. Furthermore, it employs a Kalman filter to dynamically update the time-varying mapping process online. This strategy effectively captures the slowly evolving inter-modality mapping relationship, thereby providing robust and computationally efficient support for the continuous adjustment of the tracker states.

Our main contributions are as follows:

- We pioneer the task of RGBT anti-UAV tracking under spatio-temporal misalignment and provide its first formal mathematical definition. By conceptualizing the uncalibrated observation process as a unified state-space formulation, we establish a rigorous theoretical foundation for robust tracking in complex, unaligned scenarios.
- Based on this formulation, we propose OCTA-SOT, a lightweight, training-free, and plug-and-play online adjustment framework. Its core Kalman-driven mechanism equips off-the-shelf trackers with real-time cross-modal calibration, dynamically resolving inter-sensor deviations without any retraining.
- Extensive experiments on the Anti-UAV300 dataset demonstrate that OCTA-SOT significantly enhances the DIMP tracker, achieving average improvements of 9.3% in AUC, 12.9% in Precision, and 11.9% in Normalized Precision across modalities, while surpassing state-of-the-art methods.

2 Related Work

In this section, we systematically summarize and analyze related work from three perspectives: anti-UAV tracking, unimodal tracking, and multi-modal tracking, followed by a discussion on the strengths and weaknesses of existing approaches.

2.1 Anti-UAV Tracking Methods

To address the challenges in anti-UAV tracking, research has evolved from foundational Siamese architectures to sophisticated integrative frameworks. Early methodologies predominantly leveraged Siamese networks for their efficient template matching: Jiang et al. [14] proposed DFSC using dual-stream semantic consistency, while SiamSTA [11] and SiamAD [6] incorporated spatio-temporal constraints and channel-spatial attention to bolster discriminative power. To capture more nuanced target details in complex environments, subsequent efforts focused on feature representation expansion. For instance, GASiam [25] applied graph attention to model topological dependencies, and SiamYOLO [10] fused detector-level features to enhance localization. Additionally, Huang et al. [12] designed a dual-semantic region proposal network to better distinguish UAVs from cluttered backgrounds. As the field progressed, Transformers and multi-module synergy garnered increasing attention for their ability to exploit global context. Notable works include ATMF [36], which merges re-detection with model ensembling, UTTracker [35] for robust temporal modeling, and STFTrack [33]. Most recently, the integration of detection and tracking has become the dominant paradigm, shifting toward adaptive reasoning. This is exemplified by the evidential reasoning in EDTC [43], the collaborative mechanisms in ADTC [19], and the exploitation of first-frame spatio-temporal consistency in FSTC-DIMP [4]. This trajectory reflects a clear transition from static template matching toward dynamic, unified tracking-by-detection systems.

Current anti-UAV trackers predominantly rely on a single modality, which often lacks the robustness required for complex scenarios. While RGBT data offers complementary strengths, spatiotemporal inconsistency has hindered its adoption in this field. Unlike prior works, this work introduces a lightweight online adjustment framework that operates stably even under misaligned conditions. By effectively leveraging RGBT data, our approach significantly enhances tracking performance in challenging environments.

2.2 Unimodal Tracking Methods

Single-modal object tracking has evolved significantly from CNN-based frameworks to Transformer architectures. In the deep learning era, Danelljan et al. [7] and Bhat et al. [1] established milestones with ATOM and DIMP by decoupling object classification from bounding box regression. To explicitly model spatial uncertainty, Danelljan et al. [8] subsequently introduced the probabilistic regression framework of PrDIMP. Meanwhile, the need to handle appearance variations and occlusions drove the development of state-aware mechanisms: KYS [2] utilized localized state vectors, LWL [3] employed meta-learning for rapid feature updates, and KeepTrack [22] introduced a long-term memory bank for redetection. Recently, Transformers have revolutionized the field through global context modeling. Ye et al. [34] pioneered OSTrack to unify feature extraction and fusion within a single backbone. Building on this, methods like ToMP [21] and RTS [23] introduced point-set attention and lightweight operations to balance accuracy and efficiency. Latest innovations further push the boundaries of temporal modeling, including STATrack [28] for memory distillation, MemTrack [41] for occlusion reasoning, and LLMTrack [18], which integrates large language models for semantic trajectory interpretation.

Despite substantial advancements, current single-modal paradigms still grapple with persistent bottlenecks [27,39]. Specifically, inadequate utilization of temporal priors renders their update strategies highly susceptible to tracking drift during long-term occlusions and rapid motion. By modeling both modalities and leveraging the reliable one to explicitly rectify the erroneous tracking state of the other via online cross-modal trajectory calibration with lightweight historical state queues, our approach effectively mitigates the above degradation, significantly enhancing the tracker's adaptability and robustness against prolonged occlusions and severe motion dynamics.

2.3 Multi-modal Tracking Methods

Multi-modal object tracking leverages complementary sensor data for all-weather robustness. Recently, Transformers have dominated the field by enabling deep cross-modal interactions. Specifically, Hui et al. [13] proposed TBSI, employing a template-bridging mechanism to transfer contextual information and reduce background clutter. To balance efficiency and robustness, Cao et al. [5] developed BAT, introducing a lightweight bidirectional adaptive fusion module. Concurrently, researchers have explored token-level synergies and representation learning: Lai et al. [16] designed MRTTrack with target-oriented token selection to suppress background noise; Wu et al. [31] facilitated multi-granularity fusion via token swapping in Un-Track; and Wu et al. [29] utilized modality-agnostic masking in DropMAE to learn generic cross-modal representations. Beyond interaction design, recent trends have shifted toward unified architectures and parameter-efficient adaptation. For instance, Ma et al. [20] pioneered UniSOT, a shared-parameter framework across diverse modalities. Instead of fully retraining complex networks, Zhang et al. [38] leveraged mixed adapters and learnable gating, while Zhu et al. [42] employed prompt learning in RDTTrack to seamlessly embed auxiliary infrared cues into pretrained single-modal trackers.

Despite these sophisticated fusion mechanisms, a critical systemic flaw remains: all aforementioned paradigms operate under the idealized premise of perfectly synchronized and spatially aligned sensors. In practical anti-UAV deployments, ubiquitous spatio-temporal misalignments severely corrupt cross-modal interactions. To bridge this gap, we propose OCTA-SOT. Rather than relying on static pre-calibration, our framework explicitly estimates and compensates for spatial offsets and temporal delays online, adaptively rectifying misalignments.

3 Method

In this section, we present the OCTA-SOT framework, a novel online cross-modal trajectory adjustment framework for RGBT anti-UAV single object tracking under spatio-temporal misalignment. We first formalize the problem, then detail the core components of our framework, and finally provide implementation specifics for integrating our framework with a standard baseline tracker.

3.1 Problem Formulation

This subsection establishes a rigorous theoretical foundation for the uncalibrated RGBT **single object tracking (SOT)** task. Specifically, we first model the inherent **spatio-temporal misalignment**, then define the **independent unimodal tracking** process, and finally construct a **cross-modal joint adjustment formulation** to dynamically resolve these inter-sensor deviations.

Spatio-Temporal Misalignment Modeling In uncalibrated RGBT tracking, spatio-temporal deviations are inherently dynamic. At frame k, the physical relationship between the target bounding box $\mathbf{b}_k^m$ and the delayed state $\mathbf{b}_{k-\Delta t}^{\bar{m}}$ is:

$$\mathbf{b}_k^m = \mathbf{H}_k \cdot \mathbf{b}_{k-\Delta t}^{\bar{m}} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{H}_k$, Δt, and $\boldsymbol{\epsilon}$ denote the spatial transformation, temporal delay, and dynamic noise, respectively. Since these variations make explicit calibration intractable, we formulate this misalignment as a learnable mapping over the historical trajectory:

$$\mathbf{b}_k^m = \boldsymbol{\Phi}(\tilde{\boldsymbol{\tau}}_k^{\bar{m}}), \tag{2}$$

where $\tilde{\boldsymbol{\tau}}_k^{\bar{m}}$ represents the augmented trajectory vector concatenating the past N states of modality $\bar{m}$. The operator $\boldsymbol{\Phi}(\cdot)$ is dynamically updated to adaptively align the target states across modalities.

Independent Unimodal Tracking Formulation We formulate the single-modal SOT as a state-updating process governed by a mapping function $\mathcal{F}(\cdot)$. Specifically, $\mathcal{F}$ leverages the current input image $\boldsymbol{I}_k^m$ to update the previous tracker state $\mathcal{T}_{k-1}^m$ into the current state $\mathcal{T}_k^m$, while simultaneously generating the tracking output $\boldsymbol{O}_k^m$:

$$(\mathcal{T}_k^m,\ \boldsymbol{O}_k^m) = \mathcal{F}\left(\mathcal{T}_{k-1}^m,\ \boldsymbol{I}_k^m\right), \quad m \in \{\text{vis}, \text{ir}\}, \tag{3}$$

where m denotes the modality. The output $\boldsymbol{O}_k^m$ provides the predicted target bounding box $\mathbf{b}_k^m = [l_x, l_y, w, h]^\top \in \mathbb{R}^4$.

Cross-Modal Joint Adjustment Formulation By integrating the spatio-temporal misalignment modeling into the single-modal SOT formulation, we propose a multi-modal joint correction framework for robust tracking. This unified process is formulated as:

$$\begin{aligned} \left(\mathcal{T}_{k,\text{raw}}^m,\ \boldsymbol{O}_{k,\text{raw}}^m\right) &= \mathcal{F}\left(\mathcal{T}_{k-1}^m,\ \boldsymbol{I}_k^m\right), \\ \left(\{\mathcal{T}_k^m\},\ \{\boldsymbol{O}_k^m\},\ \boldsymbol{\theta}_k\right) &= \mathcal{H}\left(\{\mathcal{T}_{k,\text{raw}}^m\},\ \{\boldsymbol{O}_{k,\text{raw}}^m\},\ \boldsymbol{\theta}_{k-1}\right), \end{aligned} \quad m \in \{\text{vis}, \text{ir}\} \tag{4}$$

where the subscript raw denotes the uncorrected intermediate results from $\mathcal{F}$. The online Joint Adjustment function $\mathcal{H}(\cdot)$ leverages these raw inputs and the global parameter set $\boldsymbol{\theta}_{k-1}$ (containing $\boldsymbol{\Phi}$) to output the updated parameters $\boldsymbol{\theta}_k$, the jointly corrected states $\{\mathcal{T}_k^m\}$, and the final tracking results $\{\boldsymbol{O}_k^m\}$.

3.2 Online Joint Adjustment Function $\mathcal{H}(\cdot)$

The online joint adjustment function $\mathcal{H}(\cdot)$ is the core module of OCTA-SOT, which refines raw tracking results of each modality by fusing temporal context and cross-modal information. It is structurally decomposed into three sequential, modular sub-functions:

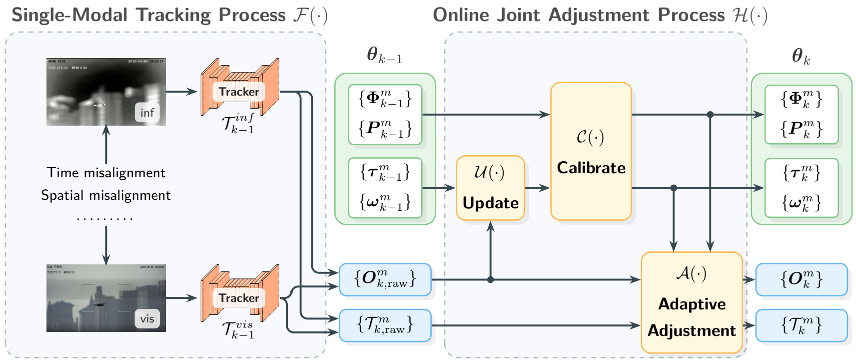

Fig. 2: Overall pipeline of the proposed OCTA-SOT.

- **History Update** $\mathcal{U}(\cdot)$: Maintains the trajectory records $\{\boldsymbol{\tau}\}$ and evaluates the reliability weights $\{\boldsymbol{\omega}\}$ for each modality.
- **Cross-Modal Calibration** $\mathcal{C}(\cdot)$: Computes the mapping parameters $\{\boldsymbol{\Phi}\}$ between modalities and their corresponding covariance $\{\mathbf{P}\}$.
- **Adaptive Adjustment** $\mathcal{A}(\cdot)$: Adjusts the unreliable tracker states $\{\mathcal{T}\}$ and generates the final robust tracking outputs $\{\boldsymbol{O}\}$.

For $m \in \{\mathrm{vis}, \mathrm{ir}\}$ (with $\bar{m}$ as its complementary modality), the unified global correction parameter $\boldsymbol{\theta}_k$ is explicitly defined as a composite set:

$$\boldsymbol{\theta}_k = \big\{\{\boldsymbol{\tau}_k^m\},\ \{\boldsymbol{\omega}_k^m\},\ \{\boldsymbol{\Phi}_{k,\bar{m}\to m}\},\ \{\mathbf{P}_{k,\bar{m}\to m}\}\big\}, \tag{5}$$

where $\boldsymbol{\tau}_k^m$ denotes the trajectory history, $\boldsymbol{\omega}_k^m$ the reliability weight sequence, $\boldsymbol{\Phi}_{k,\bar{m}\to m}$ the cross-modal mapping weight from $\bar{m}$ to m, and $\mathbf{P}_{k,\bar{m}\to m}$ the covariance matrix of the mapping weight.

As illustrated in Fig. 2, $\mathcal{H}(\cdot)$ executes in a cascaded pipeline: $\mathcal{U}(\cdot)$ first updates $\{\boldsymbol{\tau}_k^m\}$ and $\{\boldsymbol{\omega}_k^m\}$ from $\boldsymbol{\theta}_{k-1}$ and $\{\boldsymbol{O}_{k,\mathrm{raw}}^m\}$; $\mathcal{C}(\cdot)$ then estimates the mapping parameters $\{\boldsymbol{\Phi}_{k,\bar{m}\to m}\}$ and their covariance $\{\mathbf{P}_{k,\bar{m}\to m}\}$ using the updated history; and $\mathcal{A}(\cdot)$ finally outputs the refined states by fusing the raw inputs with the full $\boldsymbol{\theta}_k$. This modular structure ensures the framework is lightweight and easily integrable.

History Update $\boldsymbol{\mathcal{U}(\cdot)}$ This module maintains the per-modality trajectory records $\boldsymbol{\tau}_k^m$ and reliability weights $\boldsymbol{\omega}_k^m$. This historical preservation is crucial to capture both spatial motion trends and temporal trustworthiness, enabling reliability-aware decision-making rather than relying on a single noisy frame. At each frame k, the function $\mathcal{U}(\cdot)$ recursively updates these sequential histories based on the current raw outputs:

$$(\boldsymbol{\tau}_k^m,\ \boldsymbol{\omega}_k^m) = \mathcal{U}\left(\boldsymbol{\tau}_{k-1}^m,\ \boldsymbol{\omega}_{k-1}^m,\ \boldsymbol{O}_{k,\mathrm{raw}}^m\right). \tag{6}$$

Specifically, $\mathcal{U}(\cdot)$ appends the current frame's extracted observations to the historical sequences:

$$\boldsymbol{\tau}_k^m = \boldsymbol{\tau}_{k-1}^m \oplus \boldsymbol{p}_{k,\mathrm{raw}}^m, \quad \boldsymbol{\omega}_k^m = \boldsymbol{\omega}_{k-1}^m \oplus \mathcal{E}\big(\boldsymbol{O}_{k,\mathrm{raw}}^m\big), \tag{7}$$

where $\boldsymbol{p}_{k,\mathrm{raw}}^m$ is the target position extracted from $\boldsymbol{O}_{k,\mathrm{raw}}^m$, and $\mathcal{E}(\cdot) \in [0,1]$ evaluates the corresponding tracking confidence.

Cross-Modal Mapping Calibration $\mathcal{C}(\cdot)$ To mitigate error propagation from degraded modalities, we calibrate mapping parameters that adaptively reflect the relative reliability between sensors. Let $\boldsymbol{\Phi}_{k,\bar{m}\to m}$ and $\mathbf{P}_{k,\bar{m}\to m}$ denote the mapping parameters and their corresponding covariance from the complementary modality $\bar{m}$ to m. We formulate the calibration operator $\mathcal{C}(\cdot)$ as:

$$(\boldsymbol{\Phi}_{k,\bar{m}\to m}, \mathbf{P}_{k,\bar{m}\to m}) = \mathcal{C}\left(\boldsymbol{\tau}_k^m,\ \boldsymbol{\omega}_k^{\bar{m}},\ \boldsymbol{\theta}_{k-1}\right). \tag{8}$$

Specifically, this function employs a direction-aware gated strategy. It first computes a cross-modal reliability score $\rho_k^{\bar{m}\to m} = \mathcal{E}_{\Phi}\left(\boldsymbol{\tau}_k^m,\ \boldsymbol{\omega}_k^{\bar{m}}\right)$ to govern the calibration logic:

$$\boldsymbol{\Phi}_{k,\bar{m}\to m} = \begin{cases} \boldsymbol{\Phi}_{k-1,\bar{m}\to m}, & \text{if } \rho_k^{\bar{m}\to m} = 1, \\ \mathcal{K}\left(\boldsymbol{\tau}_k^m,\ \boldsymbol{\tau}_k^{\bar{m}},\ \boldsymbol{\Phi}_{k-1,\bar{m}\to m}\right), & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{K}(\cdot)$ denotes the Kalman filter-based update function (detailed in Sec. 3.3). This design preserves historical mappings when the current alignment is stable and triggers Kalman-based refinement otherwise to ensure temporal robustness.

Adaptive State Adjustment $\mathcal{A}(\cdot)$ Raw tracking outputs can be unreliable under occlusion or sensor noise. To prevent propagating these erroneous predictions, we introduce an adaptive adjustment module that rectifies single-modal results based on their confidence $s_k^m = \mathcal{E}_{\mathcal{A}}\left(\boldsymbol{O}_{k,\mathrm{raw}}^m\right)$. The final adjusted state and tracking output are generated by the operator $\mathcal{A}(\cdot)$:

$$(\mathcal{T}_k^m,\ \boldsymbol{O}_k^m) = \mathcal{A}\left(\mathcal{T}_{k,\mathrm{raw}}^m,\ \boldsymbol{O}_{k,\mathrm{raw}}^m,\ \boldsymbol{\theta}_k\right). \tag{10}$$

Internally, $\mathcal{A}(\cdot)$ acts as a confidence-gated switch to apply the adjustment function $\mathcal{G}(\cdot)$ only when the raw state is identified as unreliable:

$$(\mathcal{T}_k^m,\ \boldsymbol{O}_k^m) = \begin{cases} \big(\mathcal{T}_{k,\mathrm{raw}}^m,\ \boldsymbol{O}_{k,\mathrm{raw}}^m\big), & \text{if } s_k^m \geq \gamma, \\ \mathcal{G}\big(\mathcal{T}_{k,\mathrm{raw}}^m,\ \boldsymbol{O}_{k,\mathrm{raw}}^m,\ \boldsymbol{\theta}_k,\ s_k^{\bar{m}}\big), & \text{otherwise.} \end{cases} \tag{11}$$

This trust-aware design ensures high-confidence results are preserved while low-confidence, unreliable outputs are collaboratively adjusted and enhanced.

3.3 Kalman Filter-Based Mapping Weight Update

The $\mathcal{K}(\cdot)$ function calibrates cross-modal mapping weights online via a canonical Kalman filter [15]. Unlike standard tracking applications, our state variables model cross-modal geometric mappings rather than target kinematics.

Prediction: Given the previous state $\boldsymbol{\Phi}_{k-1,\bar{m}\to m}$ and covariance $\mathbf{P}_{k-1,\bar{m}\to m}$, the filter first predicts the current state and its uncertainty:

$$\hat{\boldsymbol{\Phi}}^{-}_{k,\bar{m}\to m} = \mathbf{F}\boldsymbol{\Phi}_{k-1,\bar{m}\to m}, \tag{12}$$

$$\mathbf{P}^{-}_{k,\bar{m}\to m} = \lambda \mathbf{F}\mathbf{P}_{k-1,\bar{m}\to m}\mathbf{F}^{\top} + \mathbf{Q}. \tag{13}$$

To align with the slow-varying property of cross-modal mapping, we deliberately set the state transition matrix $\mathbf{F} = \mathbf{I}$, enforcing that geometric differences between sensors change only gradually. The forgetting factor $\lambda = 0.999$ induces mild covariance decay. The process noise $\mathbf{Q} = \mathrm{diag}(10^{-4}, \dots, 10^{-4}, 1)$ is heavily regularized: small variance values constrain abrupt changes in slope coefficients, while a larger variance for the bias term accommodates slow positional drift between cameras.

Kalman Gain Computation: Next, the Kalman gain is computed to optimally weight the incoming observations from the base trackers:

$$\mathbf{K}_{k,\bar{m}\to m} = \frac{\mathbf{P}^{-}_{k,\bar{m}\to m}\boldsymbol{\tau}_{k,\bar{m}}}{\boldsymbol{\tau}^{\top}_{k,\bar{m}}\mathbf{P}^{-}_{k,\bar{m}\to m}\boldsymbol{\tau}_{k,\bar{m}} + R}. \tag{14}$$

We initialize the measurement noise variance as $R = 10^{6}$. During the initial convergence phase, raw tracking outputs are often noisy. A large R heavily penalizes these early observations, prioritizing model-based a priori predictions and preventing the filter from being misled by spurious early-frame signals.

State and Covariance Update : Finally, the mapping parameters and covariance are updated using the residual between the raw observation $y^{\mathrm{raw}}_{k,m}$ and the cross-modal prediction:

$$\boldsymbol{\Phi}^{\star}_{k,\bar{m}\to m} = \hat{\boldsymbol{\Phi}}^{-}_{k,\bar{m}\to m} + \mathbf{K}_{k,\bar{m}\to m}\big(y^{\mathrm{raw}}_{k,m} - \boldsymbol{\tau}^{\top}_{k,\bar{m}}\hat{\boldsymbol{\Phi}}^{-}_{k,\bar{m}\to m}\big), \tag{15}$$

$$\mathbf{P}_{k,\bar{m}\to m} = \big(\mathbf{I} - \mathbf{K}_{k,\bar{m}\to m}\boldsymbol{\tau}^{\top}_{k,\bar{m}}\big)\,\mathbf{P}^{-}_{k,\bar{m}\to m}. \tag{16}$$

Scaling the residual by the Kalman gain in Eq. (15) ensures that updates remain strictly proportional to uncertainty. Furthermore, to preserve positive semi-definiteness and prevent filter divergence during long-sequence tracking, we explicitly enforce posterior covariance symmetry via $\mathbf{P}_k \leftarrow \frac{1}{2}(\mathbf{P}_k + \mathbf{P}_k^{\top})$ immediately after Eq. (16).

3.4 Implementation Details on the Base Tracker

To instantiate the proposed multi-modal joint correction framework, we present a concrete implementation tailored for the DIMP [1] tracker. We first extend the base tracker function $\mathcal{F}(\cdot)$ by integrating a lightweight camera motion estimator. Specifically, we compute the inter-frame camera offset $(\Delta x, \Delta y)$ using template matching via normalized cross-correlation. This offset compensates DIMP's local search region, which is practically necessary because rapid camera motion exacerbates temporal misalignment and causes the linear assumption of our Kalman filter to break down. We then define the concrete mathematical forms of the correction functions via two complementary configuration schemes.

Scheme 1: Handcrafted Settings for General Deployment This scheme uses simple, DIMP-adaptive empirical rules requiring no per-sequence tuning. We adopt the maximum classification score from DIMP's output for confidence evaluation, setting $\mathcal{E}(\boldsymbol{O}^m_{k,\text{raw}}) = 1$ if $\max(s^m_{k,\text{cls}}) > 0.25$, and 0 otherwise. The cross-modal reliability $\mathcal{E}_{\Phi}(\cdot)$ checks if the current frame of the complementary modality is fully confident ($\omega^{\bar{m}}_{k,\text{last}} = 1$). The adjustment gating $\mathcal{E}_{\mathcal{A}}(\cdot) = \frac{1}{T}\sum_{t=k-L+1}^{k}\omega^m_t$. Furthermore, the parameter γ in Eq. (11) is set to 0.3.

The adjustment function $\mathcal{G}(\cdot)$ adaptively refines the tracker state and bounding box based on a gating threshold $\delta = 0.7$:

$$\mathcal{G}\big(\mathcal{T}^m_{k,\text{raw}}, \boldsymbol{O}^m_{k,\text{raw}}, \boldsymbol{\theta}_k, s^{\bar{m}}_k\big) = \begin{cases} \big(\mathcal{T}^m_k(\text{scale-up}), \mathbf{b}^m_{k,\text{raw}}\big), & s^{\bar{m}}_k < \delta, \\ \big(\mathcal{T}^m_k(\text{adjustment}), \boldsymbol{\phi}^\top_{k,\bar{m}}\boldsymbol{\Phi}_{k,\bar{m}\to m}\big), & s^{\bar{m}}_k \geq \delta. \end{cases} \tag{17}$$

Here, if the complementary modality is also unreliable ($s^{\bar{m}}_k < \delta$), we scale up the search region size by a factor of 1.1 to re-capture the target, denoted as $\mathcal{T}^m_k$(scale-up). If it is reliable, we update the state to $\mathcal{T}^m_k$(adjustment) and replace the bounding box with the cross-modal Kalman projection.

Scheme 2: Bayesian Optimization-Driven Settings. To maximize data adaptivity while preserving structural interpretability, Scheme 2 maintains the exact functional forms identical to the handcrafted settings in Scheme 1. In this configuration, we select seven critical dimensions for Parallel Bayesian Optimization (PBO) [26], specifically targeting the threshold parameters and the internal coefficients of the Kalman filter [15]. We employ a Gaussian Process (GP) [24] regression surrogate model with a Radial Basis Function (RBF) kernel, using Expected Improvement (EI) with an exploration parameter $\xi = 0.01$ as the acquisition function. Starting with 10 random initializations, a batch synchronous parallel strategy evaluates 10 candidates per iteration over 40 iterations. After a total of 410 tracker evaluations, the configuration yielding the highest performance is utilized for all subsequent experiments.

4 Experiments

This section evaluates the proposed OCTA-SOT. We first introduce the experimental setup, followed by extensive comparative experiments against state-of-the-art trackers, and finally validate core components through ablation studies.

4.1 Experimental Setup

We validate our approach on the Anti-UAV300 [14] dataset against representative single-modal (ATOM [7], DIMP [1], PrDIMP [8], OSTrack [34], SiamDT [12]) and state-of-the-art RGBT (TBSI [13], BAT [5], STTrack [32]) trackers. Notably, SiamDT operates as a global tracking method observing the entire image, whereas all other baselines are local trackers confined to limited search regions.

To ensure fair evaluation under unaligned conditions, single-modal baselines are trained and tested on the RGB and thermal sequences separately. Conversely, existing RGBT methods inherently assume perfect spatial alignment and typically output a single fused bounding box. To adapt them to our unaligned setting, we train and evaluate each RGBT tracker twice in two completely independent runs. In both runs, the network takes both RGB and thermal frames as inputs; however, one run is supervised and evaluated exclusively using the thermal ground truth (GT), while the parallel run strictly utilizes the visible GT.

4.2 Dataset Configuration and Evaluation Metrics

Dataset Processing. The Anti-UAV300 [14] dataset comprises 300 temporally synchronized yet spatially unaligned RGBT video pairs. Under the official split, we temporally divide the test videos into multiple sub-sequences. To comply with the standard SOT paradigm, each sub-sequence strictly begins with a valid ground-truth (GT) bounding box for tracker initialization. Additionally, we guarantee the target is never simultaneously absent in both modalities throughout any sub-sequence, thereby ensuring continuous cross-modal availability and preventing invalid evaluation penalties. Finally, both RGB and thermal frames are uniformly resized to 540×960.
Evaluation Metrics. We employ three standard metrics: (1) **Success (AUC)** [30], the area under the success plot measuring bounding box overlap; (2) **Precision (P)** [30], the percentage of frames with center location errors within 20 pixels; and (3) **Normalized Precision ($P_{\mathbf{Norm}}$)** [9], which normalizes center distance errors by target size. These metrics collectively quantify localization accuracy and stability under the uncalibrated conditions of Anti-UAV300.

4.3 Comparative Experiments

Quantitative Analysis. As shown in Table 1, OCTA-SOT achieves state-of-the-art overall performance across all metrics. Additionally, our method is not highly sensitive to hyperparameters, as evidenced by the marginal difference between the Handcrafted and PBO-Optimized variants (72.23% vs. 72.21% Mean AUC). Among the baselines, SiamDT demonstrates strong accuracy (71.45% Mean AUC) because its global search perspective effectively mitigates boundary-out drift; however, this global mechanism inevitably increases computational costs and susceptibility to visually similar distractors. Furthermore, while existing multi-modal RGBT trackers (e.g., TBSI, BAT) perform well on RGB data, they suffer severe degradation on Thermal sequences (64.51%–65.03% AUC), confirming that their rigid feature fusion fails under unaligned conditions. Conversely, our trajectory-level correction explicitly bridges this spatio-temporal gap, sustaining high accuracy across both modalities and delivering massive absolute improvements of **9.3% in Mean AUC** and **12.9% in Mean Precision** over our base tracker (DIMP).

Table 1: Quantitative comparison on the Anti-UAV300 dataset.

method	Red-Green-Blue			Thermal			Mean		
	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}
ATOM [7](CVPR2019)	56.31	74.04	71.10	58.19	76.24	74.43	57.25	75.14	72.76
DIMP [1](ICCV2019)	63.73	83.08	80.36	62.17	81.02	79.24	62.95	82.05	79.80
PrDIMP [8](CVPR2020)	63.21	82.83	80.28	65.40	86.45	83.71	64.31	84.64	81.99
OSTrack [34](ECCV2022)	69.71	92.56	88.84	62.81	83.15	80.72	66.26	87.86	84.78
SiamDT [12](TPAMI2024)	72.54	95.80	91.74	70.36	93.46	91.10	71.45	94.63	91.42
TBSI [13](CVPR2023)	72.52	95.08	91.84	64.51	84.96	82.69	68.51	90.02	87.26
BAT [5](AAAI2024)	72.68	95.02	91.32	64.62	84.99	82.38	68.19	88.90	86.30
STTrack [32](AAAI2025)	71.28	93.11	89.86	65.10	84.70	82.74	68.40	90.30	87.21
OCTA-SOT (Handcrafted)	**73.01**	**95.96**	**92.13**	71.44	93.96	91.23	**72.23**	**94.96**	**91.68**
OCTA-SOT (PBO-Optimize)	72.83	95.61	91.88	**71.60**	**94.35**	**91.29**	72.21	94.98	91.59

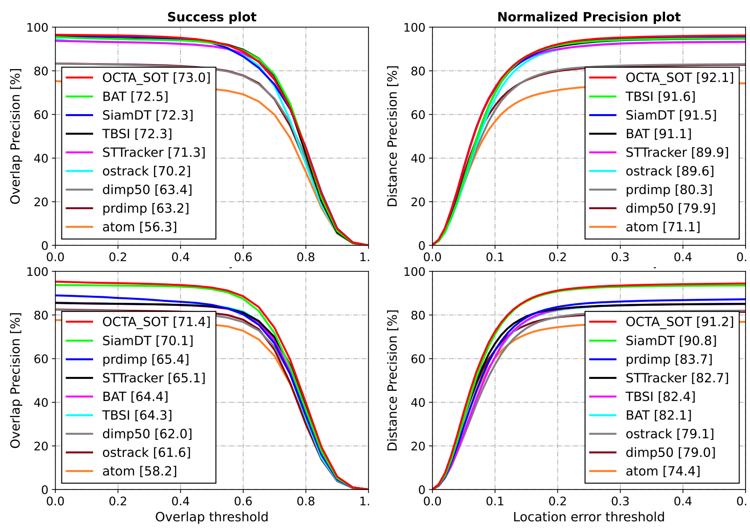

Fig. 3: Success and Normalized Precision plots for RGB (top) and Thermal (bottom) tracking modalities on the Anti-UAV300 dataset.

The robust superiority of our framework is further corroborated by the Success and Normalized Precision plots shown in Fig. 3. Across all overlap and distance error thresholds, the curves of OCTA-SOT consistently remain at the top, indicating that our method not only tracks the target successfully over long durations but also maintains highly precise center localization despite severe cross-sensor misalignment.

Qualitative Analysis. We demonstrate OCTA-SOT's efficacy through bounding box visualizations in unaligned scenarios (Fig. 4) and continuous trajectory analysis on a complex infrared sequence (Fig. 5).

In representative daytime sequences (Fig. 4, top), target scale variations and cluttered backgrounds cause conventional trackers to drift; conversely, OCTA-SOT maintains precise localization through active calibration. At nighttime (bottom), severe visible degradation necessitates reliance on thermal imaging.

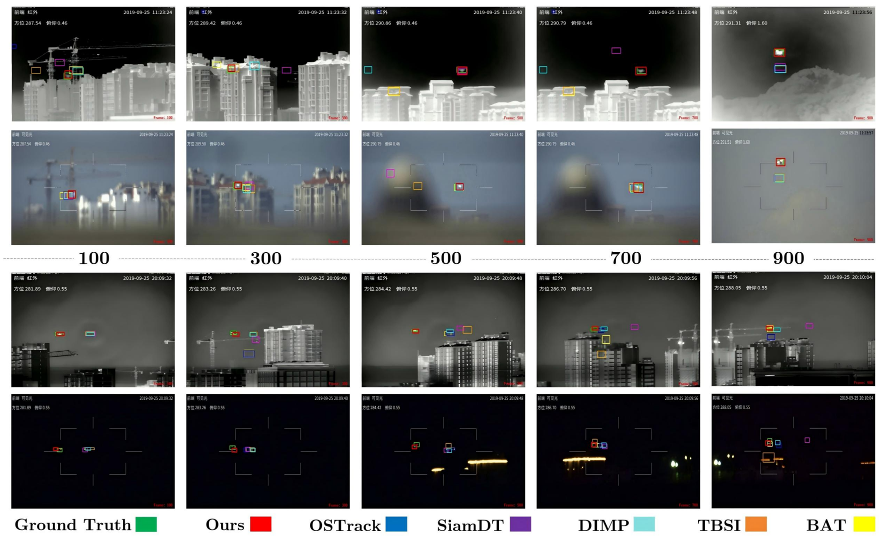

Fig. 4: Qualitative comparison on the Anti-UAV300 dataset. Top: A daytime scenario with dramatic target appearance variations. Bottom: A nighttime scenario where severe visible degradation forces reliance on the unaligned thermal modality.

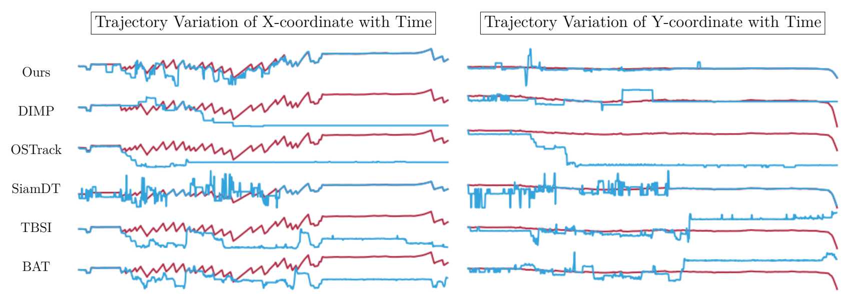

Fig. 5: Visual comparison of temporal trajectory variations in the thermal modality: red and blue lines indicate GT and predicted trajectories, respectively.

However, existing RGBT trackers fail to fully exploit multi-modal data under spatio-temporal unaligned conditions, resulting in suboptimal performance. By contrast, OCTA-SOT models spatial deviations via online Kalman-based mapping and projects reliable thermal states into the visible view, ensuring robust all-weather tracking.

Furthermore, the temporal trajectory curves in Fig. 5 confirm OCTA-SOT's superior stability. While most baselines (e.g., DIMP, TBSI, and BAT) diverge early or midway, SiamDT manages to maintain tracking for a longer duration. This is primarily due to its global search mechanism, which effectively prevents early out-of-bounds drift. However, this same global perspective accounts for its severe instability: by observing the entire image, the tracker becomes highly susceptible to visually similar distractors in the background, resulting in frequent, large-amplitude spikes and fluctuations. In contrast, OCTA-SOT follows the target trajectory smoothly, proving its superior robustness in complex, unaligned environments.

4.4 Ablation Study

We systematically evaluate our components on the Anti-UAV300 dataset starting from the DIMP baseline ($\mathcal{F}$). As reported in Table 2, adding lens motion estimation ($\mathcal{F}^*$) improves Mean AUC from 62.95 to 67.57 by keeping the target centered during rapid UAV panning. Incorporating the handcrafted joint correction module ($\mathcal{F} + \mathcal{H}_h$) alone provides a substantial +6.14% AUC boost, demonstrating that cross-modal state borrowing effectively rescues single-modality tracking failures. The most significant synergy occurs when combining both ($\mathcal{F}^* + \mathcal{H}_h$), pushing the Mean AUC to 72.23. The stabilized trajectories from $\mathcal{F}^*$ optimally satisfy the linear assumption of the Kalman filter within $\mathcal{H}$, maximizing cross-modal alignment accuracy. Finally, the handcrafted variant $\mathcal{H}_h$ performs competitively with the BO-optimized $\mathcal{H}_{pbo}$ (72.23 v.s. 72.21 Mean AUC), proving the framework's inherent robustness to hyperparameter variations and its readiness for zero-cost plug-and-play deployment.

Table 2: Ablation study on Anti-UAV300.

	Red-Green-Blue			Thermal			Mean		
	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}
$\mathcal{F}$	63.73	83.08	80.36	62.17	81.02	79.24	62.95	82.05	79.80
$\mathcal{F}^*$	67.58	88.82	85.54	67.57	88.55	86.43	67.57	88.69	85.99
$\mathcal{F} + \mathcal{H}_h$	69.87	91.62	88.16	68.30	89.50	87.02	69.09	90.56	87.59
$\mathcal{F}^* + \mathcal{H}_h$	**73.01**	**95.96**	**92.13**	71.44	93.96	91.23	**72.23**	**94.96**	**91.68**
$\mathcal{F} + \mathcal{H}_{pbo}$	70.65	92.70	89.10	68.45	89.74	87.02	69.55	91.22	88.06
$\mathcal{F}^* + \mathcal{H}_{pbo}$	72.83	95.61	91.88	**71.60**	**94.35**	**91.29**	72.21	94.98	91.59

5 Conclusion and Future Work

In this paper, we address the critical challenge of spatio-temporal misalignment in RGBT anti-UAV tracking. By formulating the uncalibrated observation process within a unified state-space model, we establish a rigorous theoretical foundation for this novel task. To tackle these complex conditions, we propose OCTA-SOT, a lightweight online adjustment framework driven by a dynamic Kalman mechanism that rectifies inter-modal deviations at the trajectory level. Extensive evaluations on the Anti-UAV300 dataset confirm that OCTA-SOT seamlessly integrates with baselines, significantly enhancing their tracking robustness and establishing a new state-of-the-art for uncalibrated environments.

Despite its efficacy, our framework has limitations. First, its reliance on explicit confidence scores restricts direct application to pure regression-based Transformers (e.g., OSTrack). Second, the linear Kalman filter requires a warm-up period, causing transient calibration delays during early frames. Future work will explore non-linear mapping (e.g., graph neural networks) for instantaneous calibration, and extend our state-space formulation to multi-view sensor clustering for large-scale defense scenarios.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62272083, 62495085, 62472066).

References

1. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6182–6191 (2019)
2. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Know your surroundings: Exploiting scene information for object tracking. In: European conference on computer vision. pp. 205–221. Springer (2020)
3. Bhat, G., Lawin, F.J., Danelljan, M., Robinson, A., Felsberg, M., Van Gool, L., Timofte, R.: Learning what to learn for video object segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 777–794. Springer (2020)
4. Bu, D., Ding, B., Tong, X., Sun, B., Sun, X., Guo, R., Su, S.: Fstc-dimp: advanced feature processing and spatio-temporal consistency for anti-uav tracking. Remote Sensing **17**(16), 2902 (2025)
5. Cao, B., Guo, J., Zhu, P., Hu, Q.: Bi-directional adapter for multimodal tracking. In: AAAI Conference on Artificial Intelligence (2024)
6. Cheng, F., Liang, Z., Peng, G., Liu, S., Li, S., Ji, M.: An anti-uav long-term tracking method with hybrid attention mechanism and hierarchical discriminator. Sensors **22**(10), 3701 (2022)
7. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4660–4669 (2019)
8. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7183–7192 (2020)
9. Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., et al.: Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision **129**(2), 439–461 (2021)
10. Fang, H., Wang, X., Liao, Z., Chang, Y., Yan, L.: A real-time anti-distractor infrared uav tracker with channel feature refinement module. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1240–1248 (2021)
11. Huang, B., Chen, J., Xu, T., Wang, Y., Jiang, S., Wang, Y., Wang, L., Li, J.: Siamsta: Spatio-temporal attention based siamese tracker for tracking uavs. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1204–1212 (2021)
12. Huang, B., Li, J., Chen, J., Wang, G., Zhao, J., Xu, T.: Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 2852–2865 (2024). `https://doi.org/10.1109/TPAMI.2023.3335338`
13. Hui, T., Xun, Z., Peng, F., Huang, J., Wei, X., Wei, X., Dai, J., Han, J., Liu, S.: Bridging search region interaction with template for rgb-t tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13630–13639 (2023)

14. Jiang, N., Wang, K., Peng, X., Yu, X., Wang, Q., Xing, J., Li, G., Guo, G., Ye, Q., Jiao, J., Zhao, J., Han, Z.: Anti-uav: A large-scale benchmark for vision-based uav tracking. IEEE Transactions on Multimedia **25**, 486–500 (2023). https://doi.org/10.1109/TMM.2021.3128047
15. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of basic Engineering **82**(1), 35–45 (1960)
16. Lai, P., Gao, D., Wang, S., Cheng, G.: Mining representative tokens via transformer-based multi-modal interaction for rgb-t tracking. Pattern Recognition p. 112162 (2025)
17. Li, C., Lu, A., Zheng, A., Tu, Z., Tang, J.: Multi-adapter rgbt tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
18. Liao, P., Yang, F., Wu, D., Yu, J., Zhu, Y., Zhao, W.: Llmtrack: Semantic multi-object tracking with multi-modal large language models. arXiv preprint arXiv:2601.06550 (2026)
19. Liu, S., Xu, T., Zhu, X.F., Wu, X.J., Kittler, J.: Learning adaptive detection and tracking collaborations with augmented uav synthesis for accurate anti-uav system. Expert Systems with Applications **282**, 127679 (2025)
20. Ma, Y., Tang, Y., Yang, W., Zhang, T., Zhou, X., Wu, F.: Unisot: A unified framework for multi-modality single object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
21. Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D.P., Yu, F., Van Gool, L.: Transforming model prediction for tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8731–8740 (2022)
22. Mayer, C., Danelljan, M., Paudel, D.P., Van Gool, L.: Learning target candidate association to keep track of what not to track. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13444–13454 (2021)
23. Paul, M., Danelljan, M., Mayer, C., Van Gool, L.: Robust visual tracking by segmentation. In: European conference on computer vision. pp. 571–588. Springer (2022)
24. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. MIT press (2006)
25. Shi, X., Zhang, Y., Shi, Z., Zhang, Y.: Gasiam: Graph attention based siamese tracker for infrared anti-uav. In: 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA). pp. 986–993. IEEE (2022)
26. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems. vol. 25 (2012)
27. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1571–1580 (2021)
28. Wang, Y., Hao, X.: Statrack: Spatio-temporal adaptive transformer with consistency-aware memory distillation for visual tracking. Knowledge-Based Systems **316**, 113360 (2025)
29. Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B.: Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14561–14571 (2023)

30. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2411–2418 (2013)
31. Wu, Z., Zheng, J., Ren, X., Vasluianu, F.A., Ma, C., Paudel, D.P., Van Gool, L., Timofte, R.: Single-model and any-modality for video object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19156–19166 (2024)
32. Xiantao, H., Ying, T., Xu, Z., Chen, Z., Zhenyu, Z., Jun, L., Bineng, Z., Jian, Y.: Exploiting multimodal spatial-temporal patterns for video object tracking. In: AAAI (2025)
33. Xie, X., Xi, J., Yang, X., Lu, R., Xia, W.: Stftrack: Spatio-temporal-focused siamese network for infrared uav tracking. Drones **7**(5), 296 (2023)
34. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European conference on computer vision. pp. 341–357. Springer (2022)
35. Yu, Q., Ma, Y., He, J., Yang, D., Zhang, T.: A unified transformer based tracker for anti-uav tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3036–3046 (2023)
36. Zhang, C., Liu, H., Xu, T., Wang, Y., Ge, S.: Atmf: Accurate tracking by multimodal fusion
37. Zhang, T., Liu, X., Zhang, Q., Han, J.: Siamcda: Complementarity-and distractor-aware rgb-t tracking based on siamese network. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1403–1417 (2021)
38. Zhang, X., Shen, Y., Huang, Z., Zhou, J., Rong, W., Xiong, Z.: Multi-modal adapter for vision-language models. arXiv preprint arXiv:2409.02958 (2024)
39. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International journal of computer vision **129**(11), 3069–3087 (2021)
40. Zhao, J., Zhang, J., Li, D., Wang, D.: Vision-based anti-uav detection and tracking. IEEE Transactions on Intelligent Transportation Systems **23**(12), 25323–25334 (2022). `https://doi.org/10.1109/TITS.2022.3177627`
41. Zhou, J., Dong, Y., Zhang, Y., Du, B.: Incorporating prior knowledge and temporal memory transformer network for satellite video object tracking. ISPRS Journal of Photogrammetry and Remote Sensing **228**, 630–647 (2025)
42. Zhu, X.F., Xu, T., Pan, Y., Gu, J., Li, X., Lu, J., Wu, X.J., Kittler, J.: Collaborating vision, depth, and thermal signals for multi-modal tracking: Dataset and algorithm. arXiv preprint arXiv:2509.24741 (2025)
43. Zhu, X.F., Xu, T., Zhao, J., Liu, J.W., Wang, K., Wang, G., Li, J., Wang, Q., Jin, L., Zhu, Z., et al.: Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-uav system. arXiv preprint arXiv:2306.15767 (2023)