

Article

CMCLTrack: Reliability-Modulated Cross-Modal Adapter and Cross-Layer Mamba Fusion for RGB-T Tracking

Pengfei Li ^{1,2,3,4}, Xiaohe Li ^{1,2} and Zide Fan ^{1,2,3,*} ¹ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100049, China² Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100089, China³ University of Chinese Academy of Sciences, Beijing 100049, China⁴ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: fanzd@aircas.ac.cn

Abstract

Single-object tracking has progressed rapidly, yet it remains fragile under low illumination, occlusion, and background clutter. RGB-Thermal (RGB-T) tracking improves robustness via modality complementarity, yet many existing trackers do not dynamically switch the dominant modality as sensing quality changes and often rely on simple late fusion at a single stage, underutilizing multi-level features across the backbone. To address these challenges, we propose CMCLTrack, a unified framework that integrates the Reliability-Modulated Cross-Modal Adapter (RMCA) and the Cross-Layer Mamba Fusion (CLMF). Specifically, RMCA performs reliability-aware bidirectional cross-modal interaction by dynamically weighting modality contributions, while CLMF efficiently aggregates complementary cues from multiple encoder layers to exploit multi-level representations. To stabilize the learning of layer-wise modality reliability, we additionally incorporate a cross-layer reliability smoothness regularization. Extensive experiments on multiple RGB-T tracking benchmarks demonstrate that CMCLTrack achieves competitive performance compared to existing state-of-the-art methods.

Keywords: RGB-T tracking; cross-modal fusion; reliability-aware interaction; cross-layer aggregation

1. Introduction

As a fundamental task in computer vision, object tracking aims to localize an arbitrary target in each video frame—a research area that has attracted extensive attention and seen significant advances over the past decades [1]. Numerous approaches and benchmarks have been proposed, leading to promising performance on RGB-based tracking. However, relying solely on the RGB modality often results in target loss or drift under challenging conditions, such as illumination changes, occlusion, and background clutter. Recently, multi-modal learning has been increasingly explored as an effective way to improve robustness against diverse real-world degradations. As a representative direction, multi-modal object tracking integrates heterogeneous sensory cues to alleviate modality-specific failures. In RGB-Thermal (RGB-T) tracking, the target state is inferred by jointly exploiting complementary cues from visible (RGB) and thermal infrared (TIR) streams. Compared with RGB images, which offer rich color and texture under favorable lighting but degrade severely in low-light scenarios, TIR images provide salient thermal contrast for heat-emitting targets while often suffering from lower spatial resolution and weaker texture details. To



Academic Editor: George A. Tsihrintzis

Received: 10 February 2026

Revised: 19 February 2026

Accepted: 20 February 2026

Published: 27 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

address the limitations of unimodal tracking, a variety of RGB-T fusion strategies have been developed to exploit the complementarity between RGB and thermal imagery, thereby improving tracking reliability across diverse and challenging environments.

Although RGB-T tracking has demonstrated clear advantages in alleviating the failures of unimodal tracking, existing methods still suffer from several critical challenges. First, in open and unconstrained environments, the reliability of different modalities is not constant [2]. RGB images are highly discriminative under favorable illumination with rich texture, whereas thermal infrared modality often becomes more reliable in low-light conditions, strong illumination interference, or severe occlusion. The dominance between RGB and thermal modalities therefore varies dynamically across time and scenes. Methods that rely on fixed dominant–auxiliary assumptions or static fusion weights may introduce noisy or misleading information when a modality degrades, consequently impairing tracking performance. This highlights the necessity of explicitly modeling modality reliability that adapts to scene variations, rather than relying on static or implicit fusion strategies. Second, from the perspective of feature representation, different layers in deep Transformer encoders exhibit strong complementarity in terms of semantic abstraction and spatial detail [3]. Shallow features are more beneficial for precise localization and boundary preservation, while deeper features are better suited for semantic discrimination and contextual reasoning. However, most existing RGB-T tracking methods perform cross-modal interaction or feature fusion at a single level—typically the final layer—thereby overlooking the potential synergy among multi-level features. This limitation restricts the model’s ability to simultaneously maintain localization accuracy and robustness in complex tracking scenarios. Finally, although recent parameter-efficient tuning approaches improve training efficiency by freezing the backbone and introducing lightweight modules, RGB-T tracking often involves more complex modality degradation patterns and diverse scene distributions. Completely freezing the backbone may hinder the model’s ability to adapt to cross-modal distribution shifts and unseen scenarios, such as aerial viewpoints or fast-moving targets, ultimately limiting generalization performance. Therefore, achieving a balance between parameter efficiency and backbone adaptability is crucial for reliable RGB-T tracking.

Another trend in RGB-T tracking is to strengthen cross-modal fusion with increasingly sophisticated interaction modules, but this often comes with non-trivial computational overhead. At the same time, exploiting multi-level or long-sequence context in Transformers is fundamentally challenged by quadratic attention complexity, which restricts efficient aggregation across layers or extended temporal contexts [4]. Inspired by recent State Space Model advances such as Mamba that enable linear-time sequence modeling [5], we adopt a Mamba-based design to perform cross-layer fusion with linear complexity, aiming to efficiently harvest hierarchical complementarities without incurring heavy computation.

Recent Transformer-based RGB-T trackers have shown strong performance by leveraging self-attention for unimodal representation learning and cross-attention for cross-modal interaction [6–8]. However, attention-based correlation estimation is often sensitive to modality degradation (e.g., low illumination in RGB or low-texture blur in TIR), which may lead to ambiguous or inappropriate interaction weights and thus harm template–search matching robustness [6,9]. Meanwhile, prompt-/adapter-based transfer frameworks further improve efficiency by reusing pretrained RGB trackers and injecting auxiliary modality cues [2,10,11]. Yet many of them implicitly assume a fixed dominant–auxiliary relation (often RGB-dominant), which is suboptimal in open scenarios where RGB/TIR reliability can swap dynamically over time and scenes [2]. These observations motivate us to explicitly model time-varying modality reliability and regulate bidirectional cross-modal prompting accordingly, rather than relying on static fusion assumptions.

To address these challenges, we propose CMCLTrack, a unified RGB-T tracking framework that couples reliability-aware cross-modal interaction with linear-complexity cross-layer fusion. Concretely, we introduce a Reliability-Modulated Cross-Modal Adapter (RMCA) that enables bidirectional information exchange between RGB and thermal streams while dynamically scaling the injection strength based on an estimated reliability signal. This design avoids committing to a fixed dominant modality and allows the tracker to emphasize the more trustworthy modality when sensing quality changes. Furthermore, to utilize hierarchical features without heavy attention-based aggregation, we develop a Cross-Layer Mamba Fusion module (CLMF) that concatenates multi-layer search tokens into a long sequence and performs cross-layer modeling with linear-time state-space scanning, producing a unified representation that captures complementary cues across depths. To stabilize the learning of layer-wise reliability, we further introduce a cross-layer reliability smoothness regularization, which discourages abrupt variations in reliability estimates between adjacent layers. Finally, instead of fully freezing the foundation backbone as in many parameter-efficient transfer settings, we adopt mild joint fine-tuning with a smaller learning rate to better adapt to cross-modal distribution shifts and unseen scenarios.

- We propose CMCLTrack, a unified RGB-T tracking framework that dynamically modulates bidirectional cross-modal interaction via a lightweight Reliability-Modulated Cross-Modal Adapter (RMCA) and efficiently aggregates multi-level representations through a linear-complexity Cross-Layer Mamba Fusion module (CLMF), thereby improving robustness under time-varying modality quality.
- We further employ a cross-layer reliability smoothness regularizer and a mild backbone fine-tuning strategy to improve training stability and generalization under challenging and unseen conditions.
- Extensive experiments on five RGB-T tracking benchmarks demonstrate the effectiveness of our method, achieving competitive performance across these datasets.

2. Related Work

2.1. Single-Object Tracking

2.1.1. Trackers Based on Appearance Modeling

With recent breakthroughs in deep learning and attention-based modeling, appearance-driven object tracking has seen substantial improvements over the past few years. Early work such as SiamFC [12] introduced a Siamese architecture to model the similarity between template and search regions, laying the foundation for modern visual tracking. Subsequent studies focused on improving feature representation by adopting stronger backbones and more effective interaction mechanisms. With the success of Transformers in computer vision, TransT [13] incorporated attention mechanisms into tracking, explicitly modeling the interaction between template and search features to enhance matching accuracy. Building upon more powerful Transformer backbones, OSTRack [1] proposed a simple yet effective one-stream tracking paradigm that unifies template and search modeling, achieving strong performance with a clean architecture. In addition, SwinTrack [14] leveraged a hierarchical Transformer backbone to capture multi-scale appearance cues, further demonstrating the effectiveness of strong appearance modeling in object tracking.

2.1.2. Trackers Combining Appearance and Temporal Context

Beyond static appearance cues, temporal context plays a crucial role in capturing appearance variations and motion patterns of targets. To improve robustness, many trackers integrate temporal information with appearance modeling. UpdateNet [15] estimates an optimal template from historical observations, representing an early attempt at adaptive template updating. STMTrack [16] introduces a memory network to explic-

itly store and fuse historical features, enabling long-term dependency modeling. More recent Transformer-based approaches explore temporal modeling in a unified manner. MixFormer [17] adopts a multi-template mechanism and jointly models multiple frames to enhance robustness against appearance changes. SeqTrack [18] formulates object tracking as a sequence prediction problem, leveraging end-to-end temporal modeling to capture long-range dependencies. Despite their effectiveness, these methods usually require additional templates or multi-frame inputs, which inevitably increase computational cost and model complexity.

2.2. Multi-Modal Tracking

2.2.1. Cross-Modal Fusion for RGB-T Tracking

Recent progress in RGB-T tracking has been closely tied to how visible and thermal cues are fused, which is commonly discussed under three paradigms: early fusion, middle fusion, and late fusion. Early-fusion methods couple the two modalities at the input stage, aiming to form a unified representation before deep feature extraction. For instance, TPF [19] adopts a pixel-level fusion strategy within a task-driven progressive learning framework to facilitate early integration of complementary cues. Middle fusion, by contrast, maintains modality-specific encoders and introduces explicit interactions at the feature level, allowing cross-modal information exchange while preserving distinctive modality characteristics. Representative examples include TBSI [19], which performs template-bridged interactions for cross-modal feature fusion; GMMT [20], which leverages generative modeling to strengthen feature-level fusion; and AINet [3], which employs a Mamba-based fusion module to enable cross-layer interactions and better exploit hierarchical representations. Late-fusion approaches postpone modality integration to the decision stage, where modality-specific predictions are aggregated to produce the final tracking result. A typical example is JMMAC [21], which combines the response maps from each modality via a linear weighting scheme.

2.2.2. Temporal Modeling in Tracking

RGB-T tracking often suffers from target appearance variations over time, making robust temporal modeling crucial for maintaining tracking stability. Early approaches mainly relied on static templates, while recent studies increasingly incorporate temporal information to improve robustness. Existing methods can be broadly categorized into three directions. The first focuses on dynamic template updating, where the target representation is adaptively updated over time, such as STMT [22], which introduces dynamic multi-modal templates. The second explores temporal feature fusion and memory mechanisms, leveraging historical features to enrich current representations, exemplified by TATrack [23]. The third direction adopts Transformer-based cross-frame interactions, modeling long-term dependencies through token propagation or unified spatiotemporal attention, as seen in CSTrack [24] and STTrack [25]. More recently, MambaVT [26] introduces state-space models for efficient long-range temporal modeling. Overall, these methods demonstrate the effectiveness of temporal modeling in RGB-T tracking, while the challenge of balancing robustness and efficiency in long-term, multi-modal scenarios remains open.

2.2.3. Parameter-Efficient Tuning

Fine-tuning is a representative transfer-learning strategy that has been investigated for decades. It typically adapts a pretrained model to a target task by optimizing the full set of network parameters with task-specific supervision. More recently, a different line of research has focused on parameter-efficient tuning (PET), which seeks to achieve effective adaptation while updating only a small set of additional parameters and keeping the backbone frozen. Initially popularized in natural language processing, approaches

such as prefix-tuning and prompt-tuning learn lightweight task-conditioned parameters, substantially lowering training and storage costs without sacrificing much transfer capability. Encouragingly, this idea has been carried over to vision. For instance, VPT [27] appends learnable prompt tokens to the input sequence and updates only the prompts while leaving the backbone untouched, demonstrating competitive performance across multiple benchmarks. In multimodal tracking, PET-style designs have also been explored to reuse strong pretrained RGB trackers with minimal modification. ProTrack [10], for example, employs a prompt-based formulation that maps multimodal observations into an RGB-compatible representation, enabling direct exploitation of pretrained RGB tracking knowledge and avoiding the need for heavy cross-modal fusion modules. Building upon this idea, ViPT [11] further designs a learnable prompt generation module that leverages the thermal infrared modality to generate visual prompts for the RGB modality, enabling effective cross-modal information injection and improving robustness under challenging conditions. Different from prompt-token-based approaches, BAT [2] adopts lightweight bi-directional adapters to perform feature prompting and bidirectional interaction between modality-specific branches, achieving efficient transfer by optimizing only the newly introduced modules and improving RGB-T tracking robustness with low parameter overhead. It is worth noting that most existing PET methods rely on freezing the backbone to ensure training stability and parameter efficiency. However, in RGB-T tracking, modality reliability varies dynamically with changes in illumination, occlusion, and thermal responses. Relying solely on a fixed backbone with lightweight auxiliary modules may therefore limit the model's ability to adapt to specific data distributions and modality degradation patterns. Motivated by this observation, our method introduces lightweight cross-modal interaction and fusion modules while not fully freezing the Transformer backbone. Instead, we adopt joint fine-tuning with a smaller learning rate for the backbone. This strategy maintains a favorable efficiency–performance trade-off, enabling effective adaptation while keeping the additional trainable parameters lightweight.

2.3. Mamba for Vision Tasks

Recently, Mamba has attracted increasing attention due to the strong scalability of State Space Models (SSMs) [5]. By introducing a selection mechanism to make structured state space sequence models such as S4 data-dependent, Mamba achieves linear-time sequence modeling while maintaining competitive capacity for long-range dependency learning, and has shown advantages over attention-based Transformers when handling long sequences. In computer vision, early attempts such as ViS4mer [28] and TranS4mer [29] demonstrate the effectiveness of SSMs for visual representation learning. More recently, Vision Mamba [30] and VMamba [31] further adapt Mamba to image modeling by designing different scanning strategies to capture interactions among image patches, enabling Mamba-style backbones to perform competitively on detection, segmentation, and tracking. Despite this progress, most existing studies primarily focus on single-modality modeling. Meanwhile, the complementarity across hierarchical feature layers has been widely observed in various vision tasks, but exploiting multi-level representations often requires attention-based aggregation with quadratic complexity. Motivated by these observations, our work mainly leverages Mamba's linear-time property to perform efficient cross-layer feature fusion, aiming to better utilize complementary cues across backbone depths without incurring heavy computational overhead.

3. Proposed Method

We propose CMCLTrack, a lightweight RGB-T tracking framework built upon a dual-stream Transformer backbone [1]. Given a template–search pair in both RGB and thermal

modalities, the backbone extracts modality-specific token representations across multiple layers. Our framework introduces three components to improve robustness under dynamically varying modality reliability and to better exploit hierarchical representations. First, we design a Reliability-Modulated Cross-Modal Adapter (RMCA) that enables explicit bidirectional interaction between modalities at each layer. In particular, a cross-modal reliability estimator (CRE) predicts layer-wise reliability weights, which modulate the information injection strength from one modality to the other. Second, to leverage complementary cues from different backbone depths, we collect intermediate fused representations from multiple layers and feed them into a Cross-Layer Mamba Fusion module (CLMF), which aggregates cross-layer token sequences with linear complexity and enhances the final representation used for prediction. Finally, the aggregated multi-modal representation is fused with the output from the last layer via gated addition, and the resulting features are fed into the prediction head to regress the target bounding box. To stabilize the learning of layer-wise reliability, we introduce an optional cross-layer reliability smoothness regularization that penalizes abrupt variations in reliability weights between adjacent layers. Overall, CMCLTrack builds upon a pre-trained backbone and achieves adaptive cross-modal interaction and efficient cross-layer aggregation for RGB-T tracking, while only requiring a mild fine-tuning of the backbone with a much smaller learning rate than the newly introduced modules. The detailed architecture of our proposed framework is illustrated in Figure 1.

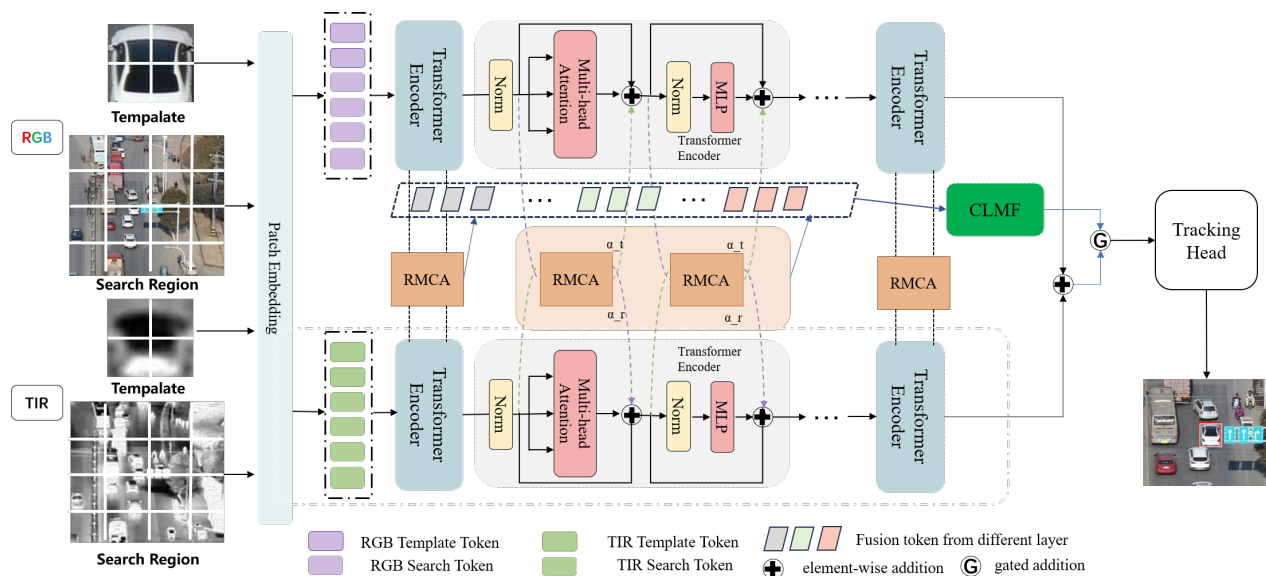


Figure 1. Overall architecture of the proposed CMCLTrack. Given a template–search pair in both RGB and thermal modalities, we first transform each modality input into token sequences via patch embedding and add positional encoding. The tokens are then processed by an N -layer dual-stream Transformer backbone, where our Reliability-Modulated Cross-Modal Adapter (RMCA) is inserted to enable explicit bidirectional interaction between modalities. To further exploit hierarchical representations, we collect fused features from multiple backbone layers and aggregate them using the Cross-Layer Mamba Fusion module (CLMF) to form a unified representation for prediction. Finally, we employ a gated addition mechanism to combine the aggregated multi-modal representation with the last layer’s features. This refined representation is then input into the tracking head to predict the target state.

3.1. Multi-Modal Tracking

Given a video V , an initial bounding box B_0 that specifies the target object Z_0 in the first frame I_{template} , and a search region I_{search} in later frames, single-modal tracking

seeks to estimate the target location over time by matching the template against the search observations. A standard tracking model T can be abstracted into a feature encoder F and a box regression head H . In transformer-based foundation trackers, I_{template} and I_{search} are first partitioned into patches and mapped to token sequences through patch embedding with positional encoding. The two token sequences are then concatenated and processed by an N -layer transformer encoder to learn joint template–search representations. After encoding, the tokens associated with the search image are selected and passed to the prediction head to infer the target state. Accordingly, the predicted bounding box is formulated as

$$B = H\left(F\left(I_{\text{template}}, I_{\text{search}}, B_0\right)\right) \quad (1)$$

where F is typically initialized from a pre-trained transformer backbone with strong representation capability.

Multi-modal tracking (MMT) extends this setting by introducing multiple synchronized and aligned video streams from different modalities, and jointly infers the target state using complementary information. For RGB-T tracking, the visible and thermal streams are captured in a frame-synchronized manner and co-registered to ensure spatial correspondence. At each time step, the tracker receives a template-search pair from both modalities, i.e., $(I_{\text{template}}^{\text{RGB}}, I_{\text{search}}^{\text{RGB}})$ and $(I_{\text{template}}^{\text{T}}, I_{\text{search}}^{\text{T}})$. The multi-modal tracking result can be expressed as

$$B = H\left(F\left(I_{\text{template}}^{\text{RGB}}, I_{\text{template}}^{\text{T}}, I_{\text{search}}^{\text{RGB}}, I_{\text{search}}^{\text{T}}, B_0\right)\right) \quad (2)$$

In our framework, we explicitly enhance cross-modal interaction and hierarchical representation learning via two key designs: the proposed RMCA module and the CLMF. Specifically, after patch and positional embedding, the RGB and thermal streams yield token sequences x_0^{RGB} and x_0^{T} . We insert RMCA into multiple encoder layers to facilitate bidirectional information exchange between the two modalities, enabling each stream to integrate modality-specific cues with complementary cues from the other modality in a layer-wise manner. Denoting the RGB and thermal tokens at layer i as x_i^{RGB} and x_i^{T} , respectively, the RMCA-enhanced dual-stream encoding process can be written as

$$\left(x_{i+1}^{\text{RGB}}, x_{i+1}^{\text{T}}\right) = F_{\text{RMCA}}^i\left(x_i^{\text{RGB}}, x_i^{\text{T}}\right), \quad i = 0, 1, \dots, N-1 \quad (3)$$

where F_{RMCA}^i denotes the i -th transformer encoder layer equipped with RMCA.

To further leverage hierarchical representations, we collect intermediate features from multiple encoder layers and aggregate them using our CLMF, producing a unified multi-level representation for robust tracking. In the last step, the fused multimodal feature is provided to the prediction head, which yields the final bounding box:

$$B = H\left(G\left(F_{\text{base}}, \text{CLMF}\left(\{x_i^{\text{RGB}}\}_{i \in \mathcal{S}}, \{x_i^{\text{T}}\}_{i \in \mathcal{S}}\right)\right)\right) \quad (4)$$

where $\mathcal{S}$ denotes the set of selected encoder layers for multi-level feature aggregation, and $\text{CLMF}(\cdot)$ represents the proposed CLMF. Moreover, $G(\cdot)$ denotes a gated fusion function that combines the baseline feature F_{base} (extracted from the last encoder layer via conventional late fusion) with the Mamba-aggregated multi-level representation produced by $\text{CLMF}(\cdot)$, yielding the final feature for prediction.

3.2. Reliability-Modulated Cross-Modal Adapter

3.2.1. Cross-Modal Reliability Estimator

Given the RGB and thermal token representations at a certain backbone layer, we estimate the global modality reliability to guide reliability-aware interaction, where the reliability score α reflects the relative feature trustworthiness (quality and discriminative

usefulness) of each modality under the current scene condition. Let $X^r \in \mathbb{R}^{B \times N \times D}$ and $X^t \in \mathbb{R}^{B \times N \times D}$ denote the RGB and thermal tokens, respectively, where each contains both template and search tokens.

Search-aware global pooling. To reflect the quality of the current tracking region, we perform global average pooling on search tokens only:

$$\mathbf{z}^r = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{x}_i^r, \quad \mathbf{z}^t = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{x}_i^t \quad (5)$$

where N_s is the number of search tokens.

Cross-modal discrepancy feature. To explicitly capture cross-modal inconsistency, we introduce a difference feature:

$$\mathbf{z}^d = |\mathbf{z}^r - \mathbf{z}^t| \quad (6)$$

The final input feature is constructed as

$$\mathbf{f} = [\mathbf{z}^r; \mathbf{z}^t; \mathbf{z}^d] \in \mathbb{R}^{3D} \quad (7)$$

Reliability prediction. We use a lightweight MLP to predict reliability logits:

$$\mathbf{l} = \text{MLP}(\mathbf{f}) \in \mathbb{R}^2 \quad (8)$$

The modality reliability scores are obtained by a temperature-scaled softmax:

$$\boldsymbol{\alpha} = \text{Softmax}\left(\frac{\mathbf{l}}{\tau}\right) = [\alpha_r, \alpha_t], \quad \alpha_r + \alpha_t = 1 \quad (9)$$

where τ is the temperature parameter. The estimated $\boldsymbol{\alpha}$ is then used to modulate the cross-modal information injection in RMCA.

3.2.2. Reliability-Modulated Cross-Modal Interaction

Following the dual-stream Transformer paradigm, we insert RMCA into each encoder layer to enable bidirectional feature prompting between the RGB and thermal branches. RMCA introduces a reliability modulation mechanism so that the strength of cross-modal injection adapts to the dynamically varying modality quality. The detailed structure of RMCA is illustrated in Figure 2.

Let $\mathbf{x}_i^r$ and $\mathbf{x}_i^t$ denote the input tokens of the i -th layer for RGB and thermal branches, respectively. Given the modality reliability weights $\boldsymbol{\alpha}_i = (\alpha_{r,i}, \alpha_{t,i})$ predicted by CRE, we modulate the injected prompts in a symmetric manner. Taking the RGB branch as an example, RMCA first generates a thermal-to-RGB prompt via a lightweight adapter $\mathcal{A}(\cdot)$:

$$\mathbf{p}_i^{t \rightarrow r} = \mathcal{A}(\mathbf{x}_i^t) \quad (10)$$

and injects it into the RGB self-attention block with reliability modulation:

$$\mathbf{x}_i^{r'} = \mathbf{x}_i^r + \mathcal{F}_{\text{Att}}(\mathbf{x}_i^r) + \alpha_{t,i} \mathbf{p}_i^{t \rightarrow r} \quad (11)$$

where $\mathcal{F}_{\text{Att}}(\cdot)$ denotes the multi-head self-attention sub-layer. Similarly, in the MLP sub-layer, RMCA performs another prompt injection to further refine the RGB representation:

$$\mathbf{x}_{i+1}^r = \mathbf{x}_i^{r'} + \mathcal{F}_{\text{MLP}}(\mathbf{x}_i^{r'}) + \alpha_{t,i} \mathcal{A}(\mathbf{x}_i^{t'}) \quad (12)$$

where $\mathbf{x}_i^{t'}$ is the intermediate thermal token after its self-attention update. The thermal branch is updated in the same way by swapping the modalities and using $\alpha_{r,i}$ for the RGB-

to-thermal injection. In this manner, RMCA encourages effective cross-modal information transfer while suppressing unreliable modality cues when the sensing quality changes over time.

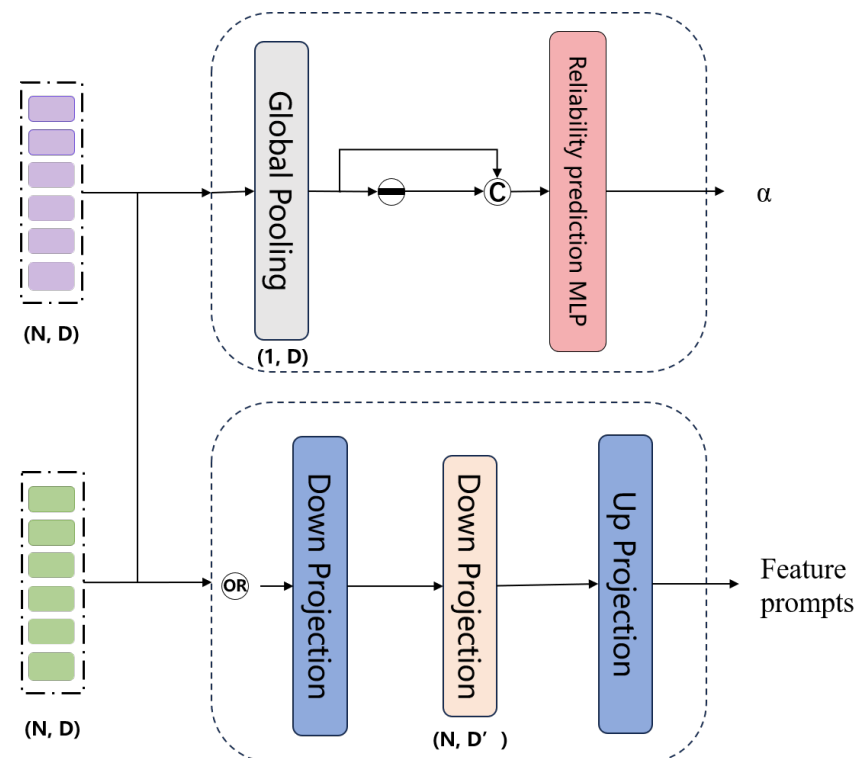


Figure 2. The detailed architecture of RMCA. The top branch estimates modality reliability weights α , where N denotes the number of tokens in each modality. The bottom branch is a bi-directional adapter that down-projects tokens to D' and up-projects them back to D to produce cross-modal feature prompts.

3.3. Cross-Layer Mamba Fusion Module

To better exploit hierarchical representations from different backbone depths, we introduce a CLMF that aggregates intermediate fused tokens collected from multiple transformer layers. Different from conventional late fusion performed only at the last layer, our design models cross-layer dependencies with linear complexity and merges the aggregated representation with the baseline feature via a warm-start gated residual.

Cross-layer token collection. Let $\mathcal{S} = \{l_1, \dots, l_K\}$ denote the indices of selected layers. At each layer $l \in \mathcal{S}$, we extract the search tokens, whose length is fixed as N_s . We first form a modality-fused representation by simple summation:

$$\mathbf{F}_l = \mathbf{X}_l^r + \mathbf{X}_l^t \in \mathbb{R}^{N_s \times D}, \quad l \in \mathcal{S} \quad (13)$$

where $\mathbf{X}_l^r$ and $\mathbf{X}_l^t$ are the RGB and thermal search-token matrices at layer l , respectively, and D is the token embedding dimension.

Sequence concatenation with level embedding. We concatenate the collected tokens along the sequence dimension to form a long token sequence:

$$\mathbf{U} = [\mathbf{F}_{l_1}; \mathbf{F}_{l_2}; \dots; \mathbf{F}_{l_K}] \in \mathbb{R}^{(KN_s) \times D} \quad (14)$$

To explicitly encode the layer-of-origin, we add a learnable level embedding $\mathbf{e}(l) \in \mathbb{R}^D$ to each token segment:

$$\tilde{\mathbf{U}}_i = \mathbf{U}_i + \mathbf{e}(l(i)) \quad (15)$$

where $l(i)$ denotes the layer index associated with the i -th token position in the concatenated sequence.

Mamba-based cross-layer fusion. We feed $\tilde{\mathbf{U}}$ into a stack of L Mamba blocks:

$$\hat{\mathbf{U}} = \text{Mamba}(\tilde{\mathbf{U}}) \in \mathbb{R}^{(KN_s) \times D} \quad (16)$$

Note that Mamba preserves the embedding dimension, i.e., the input and output share the same token dimensionality D .

Reshaping and cross-layer aggregation. We reshape $\hat{\mathbf{U}}$ back into K segments:

$$\hat{\mathbf{U}} \rightarrow \{\hat{\mathbf{f}}_{l_1}, \dots, \hat{\mathbf{f}}_{l_K}\}, \quad \hat{\mathbf{f}}_l \in \mathbb{R}^{N_s \times D} \quad (17)$$

Then we aggregate the cross-layer outputs by simple averaging to obtain a single fused search-token representation:

$$\mathbf{F}_{\text{mamba}} = \frac{1}{K} \sum_{k=1}^K \hat{\mathbf{f}}_{l_k} \in \mathbb{R}^{N_s \times D} \quad (18)$$

This aggregation summarizes complementary cues across depths while keeping the token length unchanged.

Let $\mathbf{F}_{\text{base}} \in \mathbb{R}^{N_s \times D}$ denote the baseline feature obtained by summing the last-layer RGB and thermal tokens. We combine the two via a warm-start gated residual:

$$\gamma = \sigma(g), \quad \mathbf{F}_{\text{final}} = (1 - \gamma)\mathbf{F}_{\text{base}} + \gamma\mathbf{F}_{\text{mamba}} \quad (19)$$

where g is a learnable scalar initialized to a negative value so that γ starts small, stabilizing training and gradually increasing the contribution of cross-layer fusion. Finally, we feed the fused *search* representation $\mathbf{F}_{\text{final}}$ into the prediction head, where classification and bounding-box regression are performed to localize the target.

3.4. Objective Loss

Although the proposed CRE predicts layer-wise modality reliability weights, the estimated reliability may fluctuate across adjacent layers due to local noise or abrupt feature changes. To stabilize the reliability learning and encourage consistent cross-layer modulation, we introduce a smoothness regularization on the reliability sequence across all encoder layers.

Let $\alpha_l = [\alpha_l^r, \alpha_l^t]$ denote the reliability weights predicted at layer l , where $\alpha_l^r + \alpha_l^t = 1$. Assuming the encoder has L layers, we define the smoothness loss as

$$\mathcal{L}_{\text{smooth}} = \frac{1}{L-1} \sum_{l=2}^L \|\alpha_l - \alpha_{l-1}\|_2^2 \quad (20)$$

This term penalizes abrupt variations in reliability weights between adjacent layers, yielding more stable reliability-aware modulation.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{trk}} + \lambda_s \mathcal{L}_{\text{smooth}} \quad (21)$$

where $\mathcal{L}_{\text{trk}}$ is the standard tracking loss. We set $\lambda_s = 10^{-3}$ in all experiments. This small weight makes $\mathcal{L}_{\text{smooth}}$ act as a mild regularizer to stabilize layer-wise reliability estimation, without overwhelming the primary tracking objective.

4. Experiments

4.1. Implementation Details

This section summarizes the training and evaluation settings used in our experiments. We insert the proposed RMCA module into every backbone layer. For CLMF, we aggregate multi-level features from layers 4, 7, 10, and 12 and then perform cross-layer fusion. The search and template patches are resized to 256×256 and 128×128 , respectively.

CMCLTrack is trained on 4 NVIDIA RTX A6000 GPUs with a total batch size of 32. We adopt different learning rates for different components: 5×10^{-6} for the backbone, 4×10^{-4} for RMCA, and 1.2×10^{-3} for CLMF. Optimization is conducted with AdamW [32] using a weight decay of 1×10^{-4} .

For training data and schedules, we train on the LasHeR [33] training split for 20 epochs, where each epoch contains 60K template–search pairs. The resulting model is directly tested on GTOT [34], RGBT210 [35], and RGBT234 [36] without additional fine-tuning. For VTUAV [37], we include its training split in the training stage and shorten the schedule to 10 epochs. Following [6], all runs are initialized from publicly released SOT pretraining weights [1].

4.2. Datasets and Evaluation Metrics

4.2.1. Datasets

- GTOT [34] is a grayscale–thermal object-tracking benchmark consisting of 50 aligned grayscale–thermal video pairs, covering diverse indoor and outdoor scenes such as offices, campus roads, playgrounds, and water areas. It is designed to capture representative tracking challenges, including occlusion, large-scale variation, fast motion, and low illumination, among others, and all sequences are annotated frame by frame by a single annotator to ensure high annotation consistency across modalities.
- RGBT210 [35] is an attribute-oriented RGB-thermal tracking benchmark comprising 210 aligned RGB-thermal video sequences. It provides detailed annotations of multiple tracking challenges and is widely used for attribute-based evaluation of RGB-T tracking algorithms.
- RGBT234 [36] is a widely adopted RGB-thermal tracking benchmark consisting of 234 well-aligned RGB-thermal video pairs with frame-level bounding box annotations. The dataset annotates 12 representative challenge attributes, such as occlusion, low illumination, thermal crossover, fast motion, scale variation, and camera motion, enabling comprehensive and stable evaluation of RGB-T trackers under diverse conditions.
- LasHeR [33] is a large-scale and high-diversity RGB-thermal tracking benchmark containing 1224 aligned RGB-thermal video pairs with more than 730,000 annotated frame pairs. It covers a much broader range of object categories, scene types, and imaging conditions than earlier datasets, and annotates 19 challenge attributes, including several real-world factors rarely considered before, making it particularly suitable for evaluating robustness and generalization in complex real-world scenarios.
- VTUAV [37] is an RGB-thermal tracking benchmark collected from unmanned aerial vehicle, where targets are typically small and distant, with large viewpoint and scale variations. Characterized by strong camera motion and rapid scene changes, VTUAV poses unique challenges for evaluating RGB-T trackers in aerial-view and long-range tracking scenarios.

4.2.2. Evaluation Metrics

Precision Rate (PR). PR measures the localization accuracy by the center location error (CLE) between the predicted box B_p and the ground-truth box B_{gt} :

$$CLE = \sqrt{(x_p - x_{gt})^2 + (y_p - y_{gt})^2} \quad (22)$$

where (x_p, y_p) and (x_{gt}, y_{gt}) are the center coordinates of B_p and B_{gt} , respectively. PR is defined as the percentage of frames whose CLE is smaller than a threshold τ :

$$PR = \frac{N_{CLE < \tau}}{N_{total}} \quad (23)$$

Success Rate (SR). SR evaluates the overlap between the predicted and ground-truth boxes using Intersection-over-Union (IoU):

$$IoU = \frac{Area(B_p \cap B_{gt})}{Area(B_p \cup B_{gt})} \quad (24)$$

For a given IoU threshold λ , SR is defined as:

$$SR = \frac{N_{IoU > \lambda}}{N_{total}} \quad (25)$$

In practice, the success plot is obtained by varying λ from 0 to 1, and the area under the curve (AUC) is reported as the final SR score.

Normalized Precision Rate (NPR). To reduce the sensitivity of PR to image resolution and target scale, NPR normalizes the CLE by the target size:

$$NCLE = \frac{CLE}{\sqrt{w_{gt}^2 + h_{gt}^2}} \quad (26)$$

where w_{gt} and h_{gt} denote the width and height of the ground-truth bounding box, respectively. NPR is then computed as the precision based on NCLE:

$$NPR = \frac{N_{NCLE < \tau_n}}{N_{total}} \quad (27)$$

where τ_n is the threshold for normalized precision.

Notably, in some RGB-T benchmarks (e.g., GTOT and RGBT234), the RGB/visible and thermal ground-truth boxes are not perfectly aligned due to sensor misregistration, which may bias the evaluation if a single set of annotations is used for both modalities. Following the common practice in RGB-T tracking and the protocols adopted in [38,39], we evaluate each modality with respect to its own ground truth and report the maximum precision/success across modalities, i.e., MPR/MSR, to ensure a fair comparison under misaligned annotations.

Maximum Precision Rate (MPR). To account for potential modality misalignment in multi-modal tracking, MPR adopts the minimum CLE among different modalities. For RGB-T tracking, let CLE_v and CLE_t denote the CLE measured on the visible and thermal modalities, respectively, where CLE_v and CLE_t are computed between the predicted box and the ground-truth box of the corresponding modality.

$$CLE_{min} = \min(CLE_v, CLE_t) \quad (28)$$

MPR is computed as the precision based on $CLE_{\min}$:

$$MPR = \frac{N_{CLE_{\min} < \tau}}{N_{\text{total}}} \quad (29)$$

Maximum Success Rate (MSR). Similarly, MSR uses the maximum IoU among modalities. Let IoU_v and IoU_t denote the IoU measured on the visible and thermal modalities, respectively:

$$IoU_{\max} = \max(IoU_v, IoU_t) \quad (30)$$

For a given threshold λ , MSR is defined as:

$$MSR = \frac{N_{IoU_{\max} > \lambda}}{N_{\text{total}}} \quad (31)$$

The MSR score is reported as the AUC of the maximum success plot by varying λ from 0 to 1.

4.3. Comparison with State-of-the-Art Trackers

CMCLTrack is benchmarked against representative SOTA trackers on five RGB-T datasets, including GTOT, RGBT210, RGBT234, LasHeR, and VTUAV. Following the official protocol of each benchmark, we report either PR/SR or MPR/MSR, and additionally provide NPR on LasHeR. The quantitative results are summarized in Table 1.

Table 1. Comparison of RGB-T tracking performance on GTOT, RGBT210, RGBT234, and LasHeR datasets. The arrows ($\uparrow$) indicate that higher values are better. Bold numbers denote the best results in each column.

Method	Source	GTOT		RGBT210		RGBT234		LasHeR		
		MPR $\uparrow$	MSR $\uparrow$	PR $\uparrow$	SR $\uparrow$	MPR $\uparrow$	MSR $\uparrow$	PR $\uparrow$	NPR $\uparrow$	SR $\uparrow$
DAPNet [40]	ACM MM 2019	88.2	70.7	–	–	76.6	53.7	43.1	38.3	31.4
DAFNet [41]	ICCVW 2019	89.1	71.6	–	–	79.6	54.4	44.8	39.0	31.1
MANet [42]	ICCVW 2019	89.4	72.4	–	–	77.7	53.9	45.5	–	32.6
mfDiMP [43]	ICCVW 2019	83.6	69.7	84.9	59.3	84.2	59.1	59.9	–	46.7
ADRNNet [15]	IJCV 2021	90.4	73.9	77.8	53.4	80.9	57.1	–	–	–
APFNet [44]	AAAI 2022	90.5	73.7	–	–	82.7	57.9	50.0	43.9	36.2
HMFT [37]	CVPR 2022	91.2	74.9	78.6	53.5	78.8	56.8	–	–	–
QAT [45]	ACM MM 2023	91.5	75.5	86.8	61.9	88.4	64.3	64.2	59.6	50.1
CMD [46]	CVPR 2023	89.2	73.4	–	–	82.4	58.4	59.0	54.6	46.4
ViPT [11]	CVPR 2023	–	–	–	–	83.5	61.7	65.1	–	52.5
TBSI [47]	CVPR 2023	–	–	85.3	62.5	87.1	63.7	69.2	65.7	55.6
STMT [22]	TCSVT 2024	–	–	83.0	59.5	86.5	63.8	67.4	63.4	53.7
CKD [48]	ACM MM 2024	93.2	77.2	88.4	65.2	90.0	67.4	73.2	69.3	58.1
BAT [2]	AAAI 2024	–	–	–	–	86.8	64.1	70.2	–	56.3
GMMT [20]	AAAI 2024	–	–	–	–	87.9	64.7	70.7	67.0	56.6
TATrack [23]	AAAI 2024	–	–	85.3	61.8	87.2	64.4	70.2	66.7	56.1
Un-Track [49]	CVPR 2024	–	–	–	–	83.7	61.8	66.7	–	53.6
OneTracker [50]	CVPR 2024	–	–	–	–	85.7	64.2	67.2	–	53.8
AINet [3]	AAAI 2025	–	–	86.8	64.1	89.1	66.8	73.0	69.0	58.2
CAFormer [51]	AAAI 2025	91.8	76.9	85.6	63.2	88.3	66.4	70.0	66.1	55.6
XTrack [52]	ICCV 2025	–	–	–	–	87.4	64.9	69.1	–	55.7
CMCLTrack	Ours	94.3	77.6	88.6	65.9	89.9	66.3	73.4	70.1	61.4

- GTOT. Our method achieves 94.3% MPR and 77.6% MSR, ranking 1st on both metrics. Compared with the strong baseline CKD, it brings consistent gains of +1.1% MPR (94.3 vs. 93.2) and +0.4% MSR (77.6 vs. 77.2), demonstrating superior accuracy and robustness on GTOT.

- **RGBT210.** On RGBT210, our method attains 88.6% PR and 65.9% SR, achieving the best performance on both metrics among all compared trackers in Table 1. Notably, it surpasses the previous best CKD by +0.2% PR (88.6 vs. 88.4) and +0.7% SR (65.9 vs. 65.2).
- **RGBT234.** As reported in Table 1, our method achieves 89.9% MPR and 66.3% MSR. Although our MPR is marginally below the best-performing CKD by only 0.1% (89.9 vs. 90.0), our method maintains competitive MSR performance, demonstrating strong robustness on this large-scale benchmark.
- **LasHeR.** On the challenging LasHeR dataset, our method achieves 73.4% PR, 70.1% NPR, and 61.4% SR, ranking 1st on all three metrics in Table 1. In particular, it outperforms the strong baseline CKD by +0.2% PR (73.4 vs. 73.2), +0.8% NPR (70.1 vs. 69.3), and +3.3% SR (61.4 vs. 58.1), highlighting its superior robustness under diverse and complex real-world conditions. In addition, Figure 3 compares the PR and SR curves of different methods on the LasHeR dataset. Our method consistently maintains superior performance across varying thresholds, reflecting its robustness and stability in diverse scenarios. Moreover, the attribute-based radar analysis in Figure 4 shows that our method achieves dominant performance on nearly all attributes, highlighting its strong adaptability to challenging conditions. These findings are in good agreement with the quantitative results reported in Table 1.

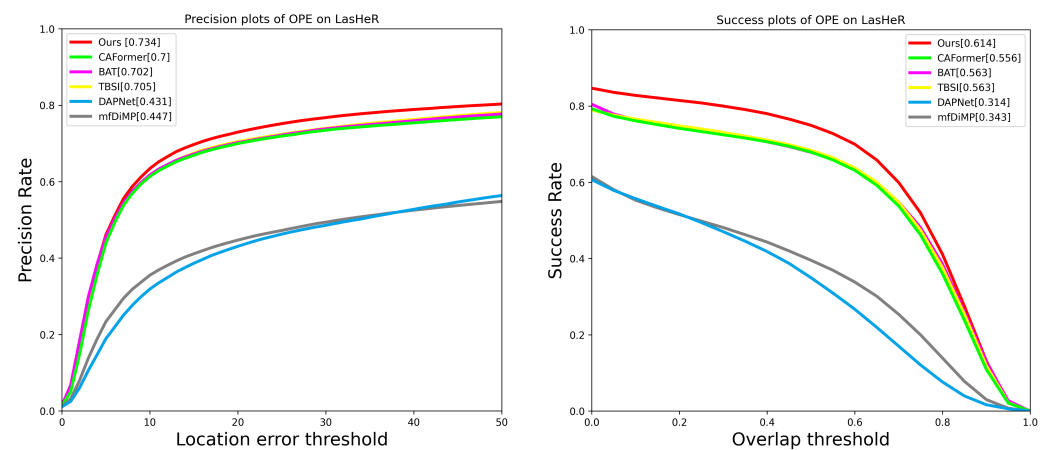


Figure 3. Evaluation curves on LasHeR dataset. The representative scores of PR/SR are presented in the legend.

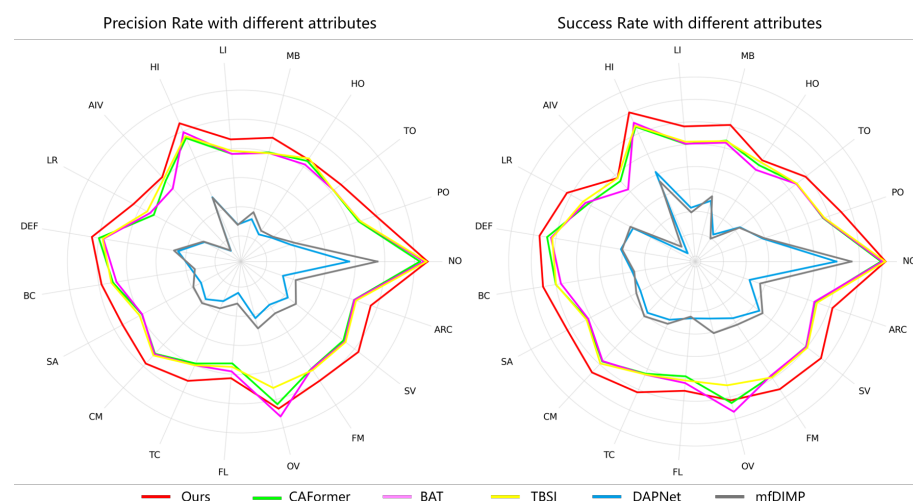


Figure 4. Attribute-based evaluation on the LasHeR dataset.

To further evaluate the scalability and generalization ability of CMCLTrack, we conduct experiments on the VTUAV benchmark, a large-scale UAV dataset characterized by complex viewpoints, rapid scale variations, and cluttered backgrounds. As reported in Table 2, on the short-term subset (VTUAV-ST), our method achieves 89.4% MPR and 78.1% MSR, yielding the best MSR among compared methods. On the long-term subset (VTUAV-LT), CMCLTrack attains 60.4% MPR and 52.5% MSR, demonstrating strong robustness under long-range temporal variations and persistent modality discrepancies. Overall, these results indicate that our approach maintains competitive performance on large-scale UAV scenarios.

Table 2. Performance on the VTUAV benchmark. The arrows (↑) indicate that higher values are better. Bold numbers denote the best results in each column.

Method	Source	VTUAV-ST		VTUAV-LT	
		MPR ↑	MSR ↑	MPR ↑	MSR ↑
DAFNet [40]	ICCVW 2019	62.0	45.8	25.3	18.8
mfDiMP [42]	ICCVW 2019	67.3	55.4	31.5	27.2
HMFT [37]	CVPR 2022	75.8	62.7	41.4	35.5
QAT [45]	ACM MM 2023	80.1	66.7	–	–
CKD [48]	CVPR 2024	90.2	77.8	–	–
AINet [3]	AAAI 2025	87.1	74.5	–	–
CAFormer [51]	AAAI 2025	88.6	76.2	–	–
CMCLTrack	Ours	89.4	78.1	60.4	52.5

4.4. Ablation Studies

4.4.1. Effect of Individual Modules

As shown in Table 3, each proposed component contributes consistently and substantially to the overall tracking performance. Compared with the baseline (52.2/42.8 on LasHeR), the only adapter variant already achieves a significant improvement (69.9/56.5), indicating that explicit cross-modal interaction is beneficial for RGB-T tracking. Building upon this, introducing RMCA further improves the performance to 70.8/57.6. This additional gain demonstrates that reliability modulation provides complementary benefits beyond standard adapter-based fusion by dynamically adjusting modality contributions. Similar trends are observed on VTUAV-ST. Further incorporating the CLMF module yields additional performance gains (LasHeR: 73.0/60.8; VTUAV-ST: 88.8/77.5), highlighting the importance of aggregating multi-level representations. By leveraging a Mamba-based design, CLMF efficiently models long-range dependencies across encoder layers while maintaining linear computational complexity. When RMCA, CLMF, and the regularization loss are jointly applied, the full model achieves the best performance (LasHeR: 73.4/61.4; VTUAV-ST: 89.4/78.1). These results confirm that RMCA and CLMF are complementary and work synergistically to enhance cross-modal robustness and hierarchical feature utilization.

Table 3. Ablation study of key components on the LasHeR and VTUAV-ST dataset. The arrows (↑) indicate that higher values are better. Bold numbers denote the best results in each column.

Method	Pretraining	LasHeR		VTUAV-ST	
		PR ↑	SR ↑	PR ↑	SR ↑
Baseline	SOT	52.2	42.8	67.3	55.2
+Only adapter	SOT	69.9	56.5	84.3	72.8
+RMCA	SOT	70.8	57.6	86.6	74.3
+CLMF	SOT	73.0	60.8	88.8	77.5
Full Model	SOT	73.4	61.4	89.4	78.1

4.4.2. Impact of Cross-Layer Fusion Depth

To determine an appropriate fusion depth for CLMF, we design ablation settings that systematically cover representative cross-layer fusion strategies used. Specifically, we evaluate: (1) last-layer-only fusion as a conventional late-fusion baseline, (2) last-three-layer fusion to emphasize deep semantic aggregation, (3) sparse multi-level fusion (layers {4, 7, 10, 12}) to balance shallow spatial details and deep context, and (4) all-layer fusion as a dense aggregation scheme. This design enables a controlled analysis of the trade-off among representation diversity, redundancy, and computational overhead, and the optimal configuration is selected based on the ablation results.

Table 4 analyzes the impact of different cross-layer fusion depths on tracking performance. Using only the last layer for fusion yields limited improvements, indicating that relying solely on high-level semantic features is insufficient for robust RGB-T tracking. Extending fusion to the last three layers (10–12) brings moderate gains on both datasets, suggesting that incorporating additional deep-layer context helps improve robustness to appearance variations. The best overall performance is achieved when fusing layers 4, 7, 10, 12, which consistently outperforms other configurations on both LasHeR and VTUAV-ST. This setting improves PR/SR to 73.4/61.4 on LasHeR and 89.4/78.1 on VTUAV-ST, demonstrating the effectiveness of aggregating complementary cues across different semantic depths. In particular, early layers preserve fine-grained spatial details that are crucial for precise localization, while intermediate and deep layers provide increasingly abstract and context-aware representations. Their joint aggregation enables the tracker to balance localization accuracy and semantic robustness in complex scenarios. Although fusing all layers achieves a slightly higher SR on LasHeR, it leads to a noticeable drop in PR and degraded performance on VTUAV-ST. This suggests that indiscriminately aggregating all layers may introduce redundant or noisy information from less informative stages, which can negatively affect localization accuracy and generalization to aerial scenarios.

Based on these observations, we adopt the 4, 7, 10, 12 configuration as a principled trade-off between performance and efficiency. This sparse yet representative layer selection effectively captures multi-level complementarities while avoiding unnecessary redundancy, making it well suited for robust and efficient RGB-T tracking.

Table 4. Ablation study on different cross-layer fusion strategies on the LasHeR and VTUAV-ST datasets.

Method	Pretraining	LasHeR		VTUAV-ST	
		PR ↑	SR ↑	PR ↑	SR ↑
Last layer only (12)	SOT	70.8	57.6	86.6	74.3
Last three layers (10–12)	SOT	71.6	59.2	87.3	75.2
Layers {4, 7, 10, 12}	SOT	73.4	61.4	89.4	78.1
All layers	SOT	72.8	61.6	88.1	76.4

4.4.3. Comparison of Cross-Layer Fusion Strategies

Table 5 compares the proposed Mamba-based fusion with a Transformer attention-based fusion counterpart. Although both strategies aggregate features from the same selected encoder layers {4, 7, 10, 12}, CLMF achieves comparable tracking performance with a slightly better accuracy while introducing substantially lower computational overhead. Specifically, the Mamba-based fusion yields a marginal performance gain over the Transformer-based variant (73.4/61.4 vs. 73.3/61.2 on LasHeR), but requires nearly half of the additional parameters (0.44 M vs. 0.79 M) and considerably fewer FLOPs (0.41 G vs. 1.07 G). This result highlights a more favorable performance–efficiency trade-off of Mamba for cross-layer aggregation.

Table 5. Performance and complexity comparison between Mamba-based and Transformer-based cross-layer fusion using encoder layers {4, 7, 10, 12}. Additional Params/FLOPs report the overhead introduced by the fusion module.

Fusion Module	LasHeR		Params (M)	FLOPs (G)
	PR $\uparrow$	SR $\uparrow$		
Mamba fusion	73.4	61.4	0.44	0.41
Transformer fusion	73.3	61.2	0.79	1.07

From an overall complexity perspective, starting from the baseline tracker, our full framework introduces only a lightweight parameter overhead (e.g., about 0.76 M additional parameters) while bringing clear performance gains.

Moreover, we introduce a gating mechanism to preserve the final-stage fused representation and to regulate the contribution of multi-layer fusion to the final prediction. By adaptively weighting the multi-layer aggregated features against the final fused feature, the gate prevents excessive interference from redundant or noisy intermediate representations and improves prediction stability.

4.4.4. Effect of Regularization Loss

Figure 5 illustrates the training and validation loss curves on the LasHeR dataset, providing insights into the optimization behavior brought by the proposed regularization loss. Although the overall performance gains introduced by this loss term are relatively moderate in quantitative evaluations, its impact on training dynamics is more pronounced. Specifically, the model equipped with the regularization loss converges significantly faster and exhibits a smoother loss decay during training. Moreover, the validation loss shows reduced oscillation and improved stability, indicating enhanced generalization and more reliable optimization. These observations suggest that the proposed regularization loss primarily contributes to stabilizing the training process and accelerating convergence, rather than merely improving the final tracking accuracy.

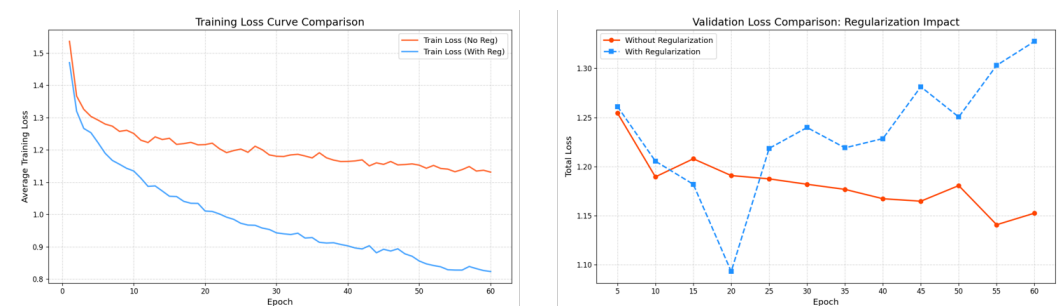


Figure 5. Training and validation loss variation during the trainings on the LasHeR dataset.

4.5. Visualization Analysis

As shown in Figure 6, we present qualitative tracking results on the LasHeR dataset. In scenarios involving severe occlusion among multiple pedestrians, our method is able to maintain accurate and stable target localization. Even when the target undergoes heavy overlap with surrounding distractors, CMCLTrack effectively avoids drift and preserves consistent tracking. This behavior indicates that the proposed cross-modal feature interaction and consistency modeling strategy can fully exploit the complementary information between RGB and thermal modalities, thereby enhancing the discriminative capability of target representations under complex occlusion conditions.

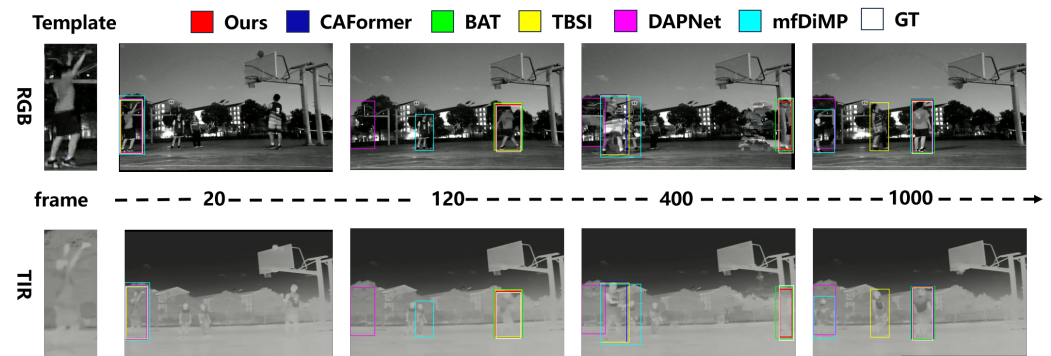


Figure 6. Visualization results on the LasHeR dataset.

Figure 7 illustrates the visualization results on the VTUAV dataset. Compared with ground-based tracking scenarios, UAVs typically suffer from large viewpoint changes, rapid scale variations, and cluttered backgrounds. Under such challenging aerial perspectives, CMCLTrack is able to produce bounding boxes that are noticeably closer to the ground-truth target compared with other competing methods. This advantage mainly stems from the effective modeling of both global and local information during cross-modal feature fusion, enabling the proposed method to better adapt to appearance variations caused by UAV viewpoints.

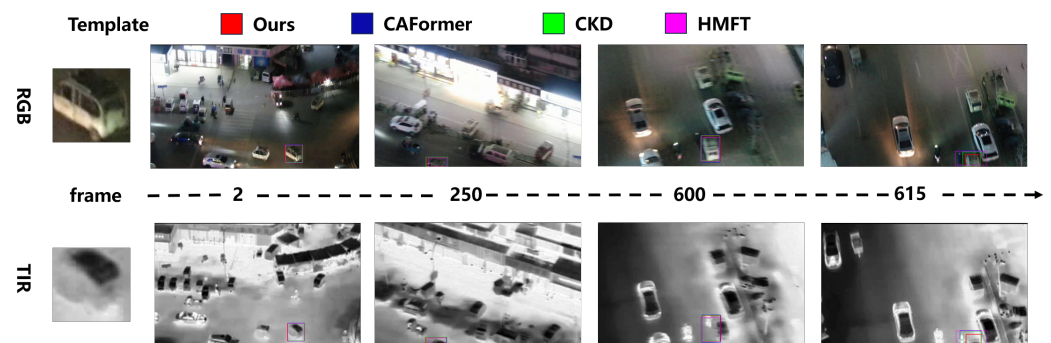


Figure 7. Visualization results on the VTUAV dataset.

5. Discussion

Beyond benchmark evaluation, our reliability-aware RGB-T tracking framework may benefit a range of real-world monitoring applications, such as industrial inspection, intelligent surveillance, and construction safety. In construction sites, safety monitoring systems often operate with heterogeneous camera sources (e.g., fixed surveillance cameras and body-worn cameras) and face challenging conditions including viewpoint changes, occlusion, motion blur, and domain shifts across devices and environments. Recent studies have investigated helmet/PPE-related monitoring under such settings, e.g., safety helmet recognition using YOLOv10 with advanced backbones on surveillance and body-worn imagery, and domain-adaptive detection for non-PPE identification to alleviate cross-domain discrepancies [53,54]. These efforts suggest that robust perception under varying sensing quality is critical in safety-sensitive scenarios. While our work focuses on RGB-T tracking, the proposed reliability modeling and fusion strategy may serve as a useful component in such systems (e.g., providing temporally consistent localization when combined with PPE detectors), and we leave task-specific integration and validation in construction scenarios as future work.

6. Conclusions

In this work, we investigate RGB-T object tracking under dynamically varying modality reliability and the need to effectively exploit hierarchical representations. To this end, we propose CMCLTrack, a unified framework that achieves reliability-aware cross-modal interaction and efficient cross-layer aggregation. Specifically, the proposed RMCA enables explicit bidirectional cross-modal interaction with reliability modulation, while the CLMF aggregates complementary cues from multiple backbone depths with linear complexity. In addition, a cross-layer reliability smoothness regularization is introduced to stabilize the learning of layer-wise reliability. Extensive experiments on multiple RGB-T benchmarks consistently demonstrate the effectiveness of our framework. Moreover, the key-component ablation studies on LasHeR and VTUAV further validate the effectiveness of the proposed modules.

Despite the promising results, several directions merit further exploration. First, reliability estimation could be enhanced by incorporating temporal consistency or uncertainty modeling to improve stability under abrupt illumination changes and severe occlusion. Second, more adaptive layer selection strategies for cross-layer fusion may further improve generalization across different datasets. Finally, extending the proposed framework to additional modalities (e.g., RGB-D) and investigating robustness under cross-dataset or cross-sensor domain shifts, such as extreme weather conditions, nighttime noise, or long-term tracking scenarios, remain interesting directions for future work.

Author Contributions: Conceptualization, Z.F.; methodology, P.L.; investigation, P.L.; data curation, X.L.; writing—original draft, P.L.; writing—review and editing, X.L.; supervision, Z.F.; funding acquisition, Z.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Laboratory Research Fund of the National Key Laboratory of Space Target Awareness, Space Engineering University [grant number STA2025ZCA0102].

Data Availability Statement: All data used in this study are publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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*; Springer: Cham, Switzerland, 2022; pp. 341–357.
2. Cao, B.; Guo, J.; Zhu, P.; Hu, Q. Bi-directional adapter for multimodal tracking. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 927–935. [[CrossRef](#)]
3. Lu, A.; Wang, W.; Li, C.; Tang, J.; Luo, B. Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 5793–5801. [[CrossRef](#)]
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
5. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, Philadelphia, PA, USA, 11 April–10 May 2024.
6. 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*, Vancouver, BC, Canada, 17–24 June 2023; pp. 13630–13639.
7. Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 19–21 June 2024; pp. 26551–26561.
8. Hou, R.; Ren, T.; Wu, G. MIRNet: A robust RGBT tracking jointly with multi-modal interaction and refinement. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, 18–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
9. Luo, Y.; Guo, X.; Dong, M.; Yu, J. Learning modality complementary features with mixed attention mechanism for RGB-T tracking. *Sensors* **2023**, *23*, 6609. [[CrossRef](#)] [[PubMed](#)]

10. Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; Song, J. Prompting for multi-modal tracking. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3492–3500.
11. Zhu, J.; Lai, S.; Chen, X.; Wang, D.; Lu, H. Visual prompt multi-modal tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9516–9526.
12. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 850–865.
13. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8126–8135.
14. Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; Ling, H. Swintrack: A simple and strong baseline for transformer tracking. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16743–16754.
15. Zhang, L.; Gonzalez-Garcia, A.; Weijer, J.V.D.; Danelljan, M.; Khan, F.S. Learning the model update for siamese trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4010–4019.
16. Fu, Z.; Liu, Q.; Fu, Z.; Wang, Y. Stmtrack: Template-free visual tracking with space-time memory networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13774–13783.
17. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 13608–13618.
18. Chen, X.; Peng, H.; Wang, D.; Lu, H.; Hu, H. Seqtrack: Sequence to sequence learning for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 20–22 June 2023; pp. 14572–14581.
19. Lu, A.; Guo, Y.; Wang, W.; Li, C.; Tang, J.; Luo, B. Breaking Shallow Limits: Task-Driven Pixel Fusion for Gap-free RGBT Tracking. *arXiv* **2025**, arXiv:2503.11247.
20. Tang, Z.; Xu, T.; Wu, X.; Zhu, X.F.; Kittler, J. Generative-based fusion mechanism for multi-modal tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5189–5197.
21. Zhang, P.; Zhao, J.; Bo, C.; Wang, D.; Lu, H.; Yang, X. Jointly modeling motion and appearance cues for robust RGB-T tracking. *IEEE Trans. Image Process.* **2021**, *30*, 3335–3347. [[CrossRef](#)] [[PubMed](#)]
22. Sun, D.; Pan, Y.; Lu, A.; Li, C.; Luo, B. Transformer RGBT tracking with spatio-temporal multimodal tokens. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 12059–12072. [[CrossRef](#)]
23. Wang, H.; Liu, X.; Li, Y.; Sun, M.; Yuan, D.; Liu, J. Temporal adaptive rgbt tracking with modality prompt. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5436–5444.
24. Feng, X.; Zhang, D.; Hu, S.; Li, X.; Wu, M.; Zhang, J.; Chen, X.; Huang, K. CSTrack: Enhancing RGB-X Tracking via Compact Spatiotemporal Features. *arXiv* **2025**, arXiv:2505.19434. [[CrossRef](#)]
25. Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; Yang, J. Exploiting multimodal spatial-temporal patterns for video object tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 3581–3589.
26. Lai, S.; Liu, C.; Zhu, J.; Kang, B.; Liu, Y.; Wang, D.; Lu, H. Mambavt: Spatio-temporal contextual modeling for robust RGB-T tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 9312–9323. [[CrossRef](#)]
27. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual prompt tuning. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 709–727.
28. Islam, M.M.; Bertasius, G. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 87–104.
29. Islam, M.M.; Hasan, M.; Athrey, K.S.; Braskich, T.; Bertasius, G. Efficient movie scene detection using state-space transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18749–18758.
30. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [[CrossRef](#)]
31. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
32. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
33. Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; Sun, D. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Trans. Image Process.* **2021**, *31*, 392–404. [[CrossRef](#)] [[PubMed](#)]
34. Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; Lin, L. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Trans. Image Process.* **2016**, *25*, 5743–5756. [[CrossRef](#)] [[PubMed](#)]

35. Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; Tang, J. Weighted sparse representation regularized graph learning for RGB-T object tracking. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1856–1864.
36. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [[CrossRef](#)]
37. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 8886–8895.
38. Hu, X.; Zhong, B.; Liang, Q.; Shi, L.; Mo, Z.; Tai, Y.; Yang, J. Adaptive perception for unified visual multi-modal object tracking. *IEEE Trans. Artif. Intell.* **2025**, *6*, 2819–2829. [[CrossRef](#)]
39. Shao, Z.; Hu, Y.; Fan, B.; Liu, H. PURA: Parameter Update-Recovery Test-Time Adaption for RGB-T Tracking. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 22089–22098.
40. Zhu, Y.; Li, C.; Luo, B.; Tang, J.; Wang, X. Dense feature aggregation and pruning for RGBT tracking. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 465–472.
41. Gao, Y.; Li, C.; Zhu, Y.; Tang, J.; He, T.; Wang, F. Deep adaptive fusion network for high performance RGBT tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October–2 November 2019; pp. 91–99.
42. Long Li, C.; Lu, A.; Hua Zheng, A.; Tu, Z.; Tang, J. Multi-adapter RGBT tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 29 October–1 November 2019; pp. 2262–2270.
43. Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; Van De Weijer, J.; Shahbaz Khan, F. Multi-modal fusion for end-to-end RGB-T tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 29 October–1 November 2019; pp. 2252–2261.
44. Xiao, Y.; Yang, M.; Li, C.; Liu, L.; Tang, J. Attribute-based progressive fusion network for rgbt tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 2831–2838.
45. Liu, L.; Li, C.; Xiao, Y.; Tang, J. Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October–3 November 2023; pp. 3129–3137.
46. Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; Han, J. Efficient RGB-T tracking via cross-modality distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 20–22 June 2023; pp. 5404–5413.
47. Li, B.; Peng, F.; Hui, T.; Wei, X.; Wei, X.; Zhang, L.; Shi, H.; Liu, S. RGB-T tracking with template-bridged search interaction and target-preserved template updating. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 634–649. [[CrossRef](#)] [[PubMed](#)]
48. Lu, A.; Zhao, J.; Li, C.; Xiao, Y.; Luo, B. Breaking modality gap in RGBT tracking: Coupled knowledge distillation. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 9291–9300.
49. 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, Seattle, WA, USA, 17–21 June 2024; pp. 19156–19166.
50. Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 19079–19091.
51. Xiao, Y.; Zhao, J.; Lu, A.; Li, C.; Yin, B.; Lin, Y.; Liu, C. Cross-modulated Attention Transformer for RGBT Tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8682–8690.
52. Tan, Y.; Wu, Z.; Fu, Y.; Zhou, Z.; Sun, G.; Zamfir, E.; Ma, C.; Paudel, D.; Van Gool, L.; Timofte, R. Xtrack: Multimodal training boosts rgb-x video object trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 5734–5744.
53. Wang, S.; Park, S.; Kim, J.; Kim, J. Safety helmet monitoring on construction sites using YOLOv10 and advanced transformer architectures with surveillance and body-worn cameras. *J. Constr. Eng. Manag.* **2025**, *151*, 04025186. [[CrossRef](#)]
54. Wang, S. Domain-adaptive faster R-CNN for non-PPE identification on construction sites from body-worn and general images. *Sci. Rep.* **2026**, *16*, 4793. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.