CADTrack: Learning Contextual Aggregation with Deformable Alignment for Robust RGBT Tracking

Hao Li[1,2], Yuhao Wang[2], Xiantao Hu[3], Wenning Hao[1*], Pingping Zhang[2*], Dong Wang[4], Huchuan Lu[2,4]

[1]College of Command and Control Engineering, Army Engineering University of PLA
[2]School of Future Technology, Dalian University of Technology
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology
[4]School of Information and Communication Engineering, Dalian University of Technology
lihao@aeu.edu.cn, 924973292@mail.dlut.edu.cn, xiantaohu@njust.edu.cn, hwnbox@aeu.edu.cn,
{zhpp,wdice,lhchuan}@dlut.edu.cn

Abstract

RGB-Thermal (RGBT) tracking aims to exploit visible and thermal infrared modalities for robust all-weather object tracking. However, existing RGBT trackers struggle to resolve modality discrepancies, which poses great challenges for robust feature representation. This limitation hinders effective cross-modal information propagation and fusion, which significantly reduces the tracking accuracy. To address this limitation, we propose a novel **C**ontextual **A**ggregation with **D**eformable Alignment framework called **CADTrack** for RGBT Tracking. To be specific, we first deploy the Mamba-based Feature Interaction (MFI) that establishes efficient feature interaction via state space models. This interaction module can operate with linear complexity, reducing computational cost and improving feature discrimination. Then, we propose the Contextual Aggregation Module (CAM) that dynamically activates backbone layers through sparse gating based on the Mixture-of-Experts (MoE). This module can encode complementary contextual information from cross-layer features. Finally, we propose the Deformable Alignment Module (DAM) to integrate deformable sampling and temporal propagation, mitigating spatial misalignment and localization drift. With the above components, our CADTrack achieves robust and accurate tracking in complex scenarios. Extensive experiments on five RGBT tracking benchmarks verify the effectiveness of our proposed method.

Code — https://github.com/IdolLab/CADTrack

Introduction

RGB-Thermal (RGBT) tracking is a fundamental task in computer vision and image processing. It estimates object states by fusing complementary information from visible (RGB) and thermal infrared (TIR) modalities. Generally, the RGB modality provides rich textures and color details under normal lighting. The TIR modality provides structural and semantic information under lighting variations. The fusion of these modalities can address limitations of single-modality trackers, such as failure in illumination changes, occlusions, or sensor noise. These advantages make RGBT tracking suitable for all-day and all-weather scenarios.

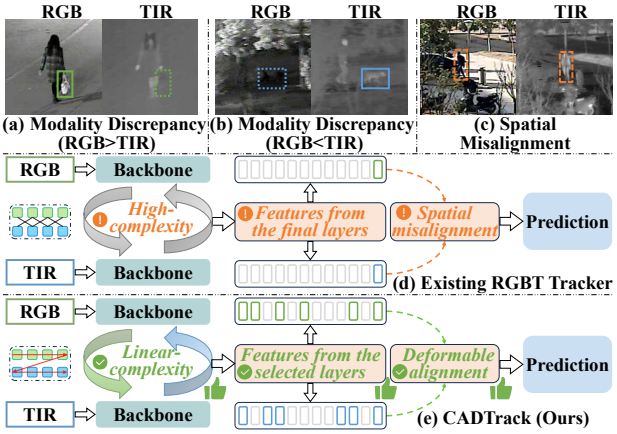

Figure 1: Comparison with different RGBT tracking paradigms. (a)-(c) The limitations of RGBT tracking include modality discrepancies and spatial misalignment. (d) Existing RGBT trackers exhibit high-complexity feature interaction, use features only from the final layers, and suffer from spatial misalignment. (e) Our framework introduces linear-complexity modality interaction, selects features from multiple layers, and employs deformable alignment.

In fact, most existing RGBT trackers focus on cross-modal fusion (Xia et al. 2024; Fan et al. 2024; Hu et al. 2025c; Tan et al. 2025a). However, they overlook modality discrepancies, as shown in Fig. 1. First, complex fusion methods (Chen and Wang 2025) achieve robust information interaction but always result in high computational costs, which hinder real-time performance. Second, prior methods (Zhu et al. 2023b, 2024b,a; Xu et al. 2024; Feng et al. 2025a; Cai, Liu, and Wang 2025; Liu et al. 2025; Guo et al. 2025) extract features solely from the final layers of backbone networks. They neglect cross-layer semantic information that could enhance occlusion handling and localization accuracy. Third, modality discrepancies often lead to spatial

*Corresponding authors.

misalignment during motion or viewpoint changes (Zhang et al. 2024). This naturally raises the question: ***Can we design a RGBT tracker that dynamically adapts to modality discrepancies while ensuring computational efficiency?***

To address the above challenges, we propose CADTrack, a novel framework that combines linear-complexity feature interaction, contextual aggregation, and deformable alignment. As shown in Fig. 2, our framework comprises three key components: Mamba-based Feature Interaction (MFI), Contextual Aggregation Module (CAM), and Deformable Alignment Module (DAM). The MFI employs State Space Models (SSMs) (Gu et al. 2021) to establish efficient feature interaction, reducing computation cost while preserving contextual coherence. Meanwhile, the CAM dynamically activates backbone layers through sparse gating based on the Mixture-of-Experts (MoE), which achieves contextual aggregation of multi-level features. By leveraging complementary information from shallow-level and deep-level features, it addresses the challenge of cross-layer feature underutilization. Finally, the DAM maintains spatial alignment through modality-specific deformable sampling and temporal propagation. Extensive experiments on five benchmarks verify the effectiveness of our method, achieving state-of-the-art performance in both accuracy and robustness.

Our main contributions are summarized as follows:

- We propose CADTrack, a novel framework combining linear-complexity feature interaction, contextual aggregation, and deformable alignment for RGBT tracking.
- We propose a Contextual Aggregation Module (CAM) that activates backbone layers through sparse gating, facilitating contextual aggregation of multi-level features.
- We propose a Deformable Alignment Module (DAM) to solve cross-modal spatial misalignment through deformable sampling and temporal propagation.
- Extensive experiments on five RGBT tracking benchmarks demonstrate that our method achieves state-of-the-art performance while maintaining real-time efficiency.

Related Work

Cross-Modal Fusion in RGBT Tracking

Recent advances have catalyzed RGBT tracking into three primary paradigms: early fusion, middle fusion and late fusion. Early fusion means fusing two modalities in the input level. For instance, TPF (Lu et al. 2025a) introduces a pixel-level fusion strategy by leveraging a task-driven progressive learning framework. Middle fusion first extracts features independently from two modalities and then fuses them at the feature level. For example, TBSI (Hui et al. 2023) employs cross-modal feature fusion through template-bridged interactions. GMMT (Tang et al. 2024) enhances feature-level fusion by leveraging generative models. AINet (Lu et al. 2025b) utilizes Mamba-based feature fusion to achieve all-layer interactions. Late fusion means aggregating tracking results from two modalities. For example, JMMAC (Zhang et al. 2021) implements late fusion by linearly combining modality-specific response maps. Additionally, methods that leverage Parameter-Efficient Fine-Tuning (PEFT), such as ProTrack (Yang et al. 2022), ViPT (Zhu et al. 2023a), BAT (Cao et al. 2024), SDSTrack (Hou et al. 2024) and OneTracker (Hong et al. 2024), enable RGB-to-RGBT transfer. However, current methods suffer from high computational cost and inadequate cross-modal fusion. Different from existing fusion methods, our method constructs a unified feature representation space that bridges heterogeneous modality characteristics with linear-complexity SSMs.

Mixture-of-Experts in RGBT Tracking

MoE routes inputs to specialized experts via a learnable gate, performing an adaptive weighted sum over selected experts (Yuksel, Wilson, and Gader 2012). For RGBT tracking, MoE has been leveraged to adapt to modality-specific variations by activating experts customized for RGB or TIR features. For example, AETrack (Zhu et al. 2025) employs multiple experts for feature extraction. MoETrack (Tang et al. 2025) utilizes decision-level experts with confidence-based routing. XTrack (Tan et al. 2025b) incorporates shared and modality-specialized experts for input routing. These advances improve robustness by adapting to modality variations and optimizing the expert selection. However, current methods ignore the adaptive fusion of multi-level features. This limits their ability to handle occlusion and scale challenges. Different from existing MoE-based methods, our method aggregates cross-layer features through sparse gating. This allows a contextual aggregation of multi-level features to enhance tracking robustness.

Temporal Modeling in RGBT Tracking

RGBT tracking addresses dynamic targets with evolving appearances, requiring robust spatiotemporal modeling. Previous methods rely on static templates, while recent methods aim to leverage temporal information. For example, STMT (Sun et al. 2024) introduces dynamic multi-modal templates. TATrack (Wang et al. 2024) integrates temporal-aware feature fusion. CSTrack (Feng et al. 2025b) unifies spatiotemporal fusion within a Vision Transformer (ViT) backbone. STTrack (Hu et al. 2025a) establishes inter-frame token propagation for long-range dependencies. MambaVT (Lai et al. 2025) pioneers a Mamba-based architecture for long-range modeling. However, current methods neglect explicit spatial misalignment. Different from existing methods, our method resolves this limitation through spatial correction and spatiotemporal feature propagation.

Methodology

In this paper, we propose CADTrack for RGBT tracking, which includes three key components: Mamba-based Feature Interaction (MFI), Contextual Aggregation Module (CAM) and Deformable Alignment Module (DAM). The overall framework of CADTrack is shown in Fig. 2. MFI facilitates efficient cross-modal feature interaction via SSMs, enabling selective feature fusion while maintaining linear complexity. CAM aggregates multi-level features through sparse gating based on MoE. DAM generates spatiotemporal alignment cues by integrating deformable sampling with temporal propagation, resolving spatial misalignment

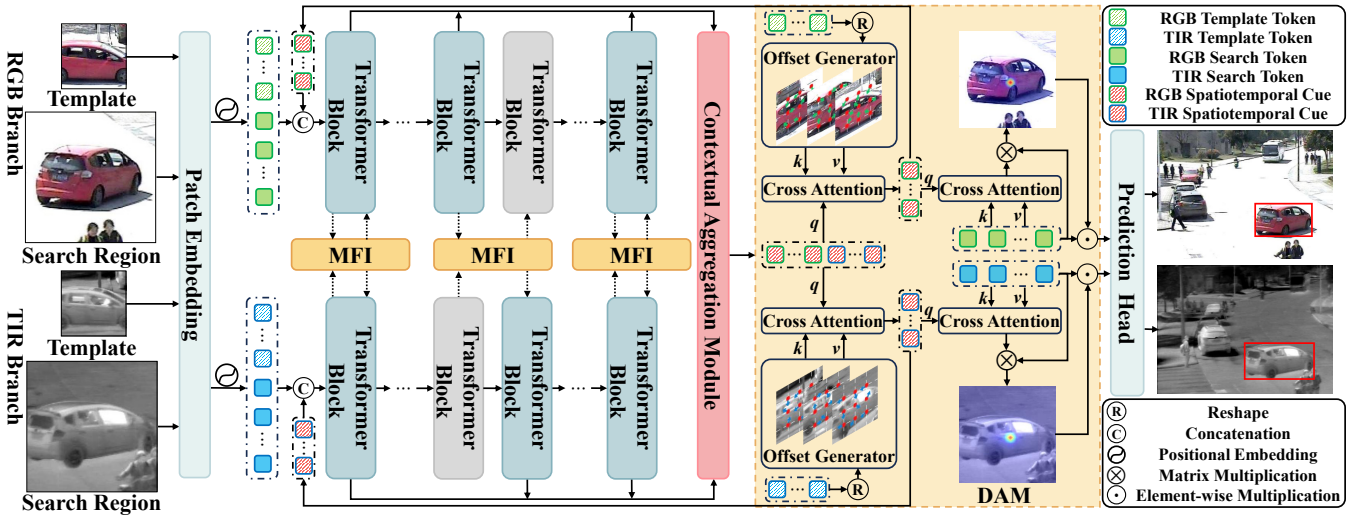

Figure 2: Overall framework of our proposed CADTrack. Firstly, input templates and search regions are tokenized with spatiotemporal alignment cues from previous frames. Then, they are fed into the backbone network with MFI for selective feature interaction. Meanwhile, CAM aggregates multi-level contextual features using modality-specific sparse gating, while DAM generates updated cues through spatial guidance for precise alignment. Finally, a prediction head is used for target localization.

between modalities. The overall framework and key components are described in the following subsections.

Overall Framework

The RGBT tracking task is initialized by a target annotation in the first image pair of RGB and TIR modalities. At time step t, we process initial templates $\mathbf{Z}_m^0 \in \mathbb{R}^{3\times H_z\times W_z}$, dynamic templates $\mathbf{Z}_m^t \in \mathbb{R}^{3\times H_z\times W_z}$, and search regions $\mathbf{S}_m^t \in \mathbb{R}^{3\times H_x\times W_x}$ for modalities $m \in \{R, T\}$. Each image is split into $P \times P$ non-overlapping patches and linearly projected into C-dimensional tokens $\mathbf{F}_m^{Z_0} \in \mathbb{R}^{N_z\times C}$, $\mathbf{F}_m^{Z_t} \in \mathbb{R}^{N_z\times C}$, $\mathbf{F}_m^{S_t} \in \mathbb{R}^{N_x\times C}$, where N_z and N_x denote the numbers of patch tokens.

Then, we construct modality-specific representations by concatenating features with spatiotemporal alignment cues $\mathbf{C}_m^t$ from previous frames:

$$\mathbf{F}_m^0 = \left[\mathbf{C}_m^t; \mathbf{F}_m^{Z_0}; \mathbf{F}_m^{Z_t}; \mathbf{F}_m^{S_t}\right]. \tag{1}$$

These tokens are processed through Transformer blocks:

$$\mathbf{F}_m^l = \mathcal{T}^l(\mathbf{F}_m^{l-1}), \tag{2}$$

where $l \in \{1, 2, \ldots, L\}$ is the layer number. Each Transformer $\mathcal{T}^l$ (Dosovitskiy et al. 2021) contains Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN).

To enable efficient feature interaction, we introduce MFI between different modalities via SSMs. For multi-level feature aggregation, we propose CAM to combine the output features from all backbone layers. Meanwhile, we propose DAM to generate spatiotemporal alignment cues $\mathbf{C}_m^{t+1}$ for subsequent frames. These cues enhance the aggregated features via cross-attention. Finally, the multi-modal features $\mathbf{H}_m^t$ are concatenated and compressed via convolutional operations to reduce channels while preserving spatial resolution. The obtained features are fed into the prediction head:

$$\mathbf{B}^t = \phi\left(\left[\mathbf{H}_R^t; \mathbf{H}_T^t\right]\right), \tag{3}$$

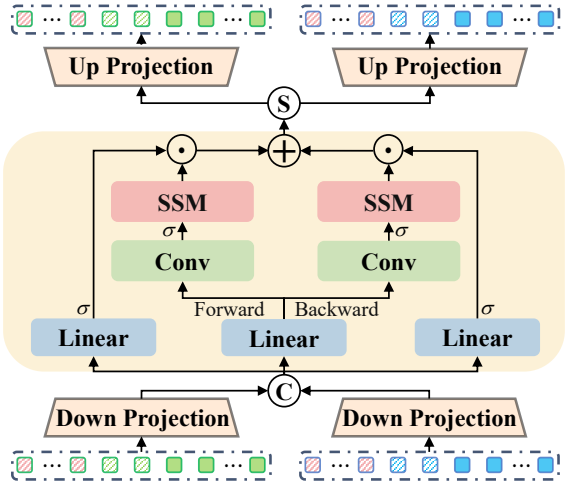

Figure 3: Details of our proposed MFI.

where ϕ is a fully convolutional network with stacked Conv-BN-ReLU layers, and $\mathbf{B}^t$ denotes the tracking result.

Mamba-based Feature Interaction

To fully bridge RGB and TIR modalities, we propose the MFI, a cross-modal interaction module leveraging Mamba. Traditional interaction methods suffer from high computational cost. In contrast, our MFI adapts to modality-specific characteristics while maintaining computational efficiency.

As shown in Fig. 3, the features $\mathbf{F}_m^l$ at the l-th layer are first projected into a shared latent space:

$$\hat{\mathbf{F}}_m^l = \mathbf{W}_m^{down}\mathbf{F}_m^l, \tag{4}$$

where $\mathbf{W}_m^{down}$ is a learnable projection matrix that performs down-sampling by reducing the channel dimension. It sig-

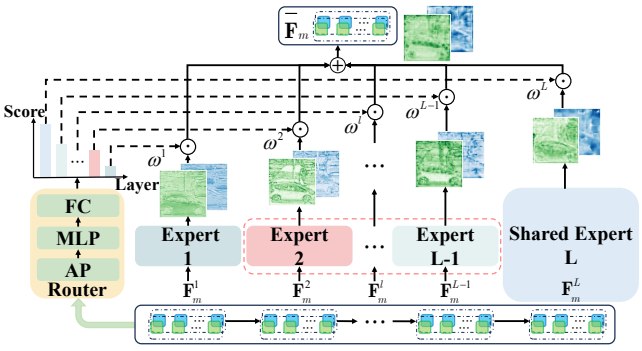

Figure 4: The structure of our proposed CAM.

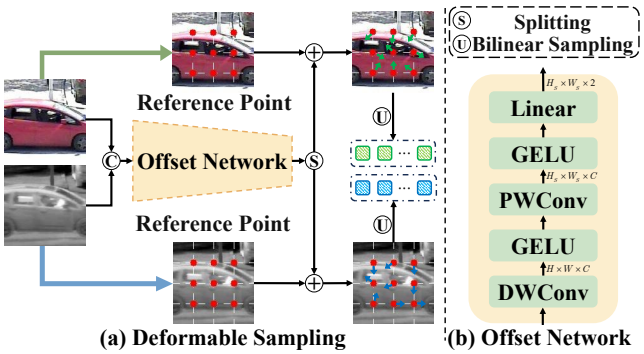

Figure 5: Deformable alignment of DAM.

nificantly decreases computational overhead while aligning feature representations across modalities.

Then, the projected features are concatenated into a unified sequence $\mathbf{F}_{RT}$, which is processed through the forward Mamba $\mathcal{M}_f$ and backward Mamba $\mathcal{M}_b$:

$$\hat{\mathbf{F}}_{RT} = \mathcal{M}_f(\mathbf{F}_{RT}) + \mathcal{M}_b(\mathbf{F}_{RT}), \tag{5}$$

$$\mathcal{M}_*(x) = \text{SSM}_*(\sigma(\mathcal{D}(\Gamma(x)))) \odot \sigma(\Gamma(x)), \tag{6}$$

where Γ is the linear projection. $\mathcal{D}$ is the convolution operation. σ is the SiLU activation (Elfwing, Uchibe, and Doya 2018). SSM_f and SSM_b are the forward and backward SSM, respectively. $\odot$ is the element-wise multiplication.

Afterwards, the features are split into $\tilde{\mathbf{F}}^l_m$, projected back to the original dimension via $\mathbf{W}^{up}_m$, and adaptively fused through residual connections:

$$\mathbf{F}^l_m = \mathbf{F}^l_m + \mathbf{W}^{up}_m \tilde{\mathbf{F}}^l_m. \tag{7}$$

This module establishes a unified feature space for RGB and TIR modalities, enabling efficient cross-modal feature interaction. Unlike previous methods, our MFI can enhance the robustness against feature degradation while maintaining the linear computation complexity.

Contextual Aggregation Module

Existing feature aggregation methods face two critical limitations. First, fixed layer selection strategies (Li et al. 2025) fail to adapt to dynamic scene variations. Second, simple aggregation strategies (Lu et al. 2025b) amplify noise and drift risks in RGBT tracking. Therefore, we propose the CAM to address scene-specific challenges and achieve robust feature aggregation. Specifically, our CAM leverages sparse activation principles to dynamically select multi-level features.

As shown in Fig. 4, our proposed CAM establishes parallel expert pools for RGB and TIR modalities. Besides, a modality-shared router $\mathcal{R}$ generates expert selection scores s_m by aggregating features from all L Transformer layers:

$$s_m = \mathcal{R}\left(\left[\mathbf{F}^1_m; \mathbf{F}^2_m; \ldots; \mathbf{F}^L_m\right]\right), \tag{8}$$

$$\mathcal{R}(x) = \mathcal{Y}(\mathcal{P}(\mathcal{A}(x))), \tag{9}$$

where $\mathcal{A}$ is the Average Pooling (AP). $\mathcal{P}$ is the Multilayer Perceptron (MLP). $\mathcal{Y}$ is the Fully Connected (FC) layer. The selected layers are then processed through expert networks:

$$\tilde{\mathbf{F}}^l_m = \mathbf{W}^l \mathbf{F}^l_m, \tag{10}$$

where $\mathbf{W}^l$ denotes modality-shared projection weights. The linear projection mitigates feature distribution gaps across different layers (Liu et al. 2023b).

Our CAM is implemented based on three key insights: (1) The shallowest expert ($l = 1$) is always activated to preserve high-frequency spatial details essential for object deformation; (2) The deepest expert ($l = L$) is selected as the shared expert to maintain semantic information during contextual aggregation; (3) The remaining experts are selected via top-k sparse activation based on router scores s_m.

The resulting modality-specific aggregation is performed over the selected expert set $\mathcal{E}_m$ using learnable weights w^l for selected layers, as follows:

$$\overline{\mathbf{F}}_m = \sum_{l \in \mathcal{E}_m} w^l \odot \tilde{\mathbf{F}}^l_m. \tag{11}$$

Unlike static feature aggregation methods, our proposed CAM enables context-aware expert selection that adapts to scene variations while maintaining computational efficiency.

Deformable Alignment Module

To address the spatial misalignment caused by modality discrepancies, we propose the DAM, which generates spatiotemporal alignment cues $\mathbf{C}^t_m \in \mathbb{R}^{N_K \times C}$ via deformation sampling and temporal propagation.

As shown in Fig. 5, we first extract patch tokens from initial template features and dynamic template features, and reshape them into spatial representations $\hat{\mathbf{F}}^{Z_0}_m \in \mathbb{R}^{H \times W \times C}$, $\hat{\mathbf{F}}^{Z_t}_m \in \mathbb{R}^{H \times W \times C}$, where $H = H_z/P$ and $W = W_z/P$. Then, these features are concatenated and processed through a convolutional mixer $\mathcal{O}$:

$$\hat{\mathbf{F}}_A = \mathcal{O}\left(\left[\hat{\mathbf{F}}^{Z_0}_m; \hat{\mathbf{F}}^{Z_t}_m\right]\right), \tag{12}$$

$$\mathcal{O}(x) = \delta(\mathcal{W}(\delta(\mathcal{Q}(x)))), \tag{13}$$

where $\mathcal{W}$ and $\mathcal{Q}$ are the Point-Wise Convolution (PWConv) and Depth-Wise Convolution (DWConv), respectively. δ is the GELU activation (Hendrycks and Gimpel 2016).

With reference points $\tilde{P} \in \mathbb{R}^{H_S \times W_S \times 2}$ from convolution centers, we predict modality-specific offsets to accommodate the spatial misalignment:

$$\Delta\tilde{P}^\tau_m = v \cdot \mathcal{G}^\tau_m(\hat{\mathbf{F}}_A), \tag{14}$$

Method	Source	Backbone	Pretraining	GTOT		RGBT210		RGBT234		LasHeR		
				MPR↑	MSR↑	PR↑	SR↑	MPR↑	MSR↑	PR↑	NPR↑	SR↑
DAPNet (Zhu et al. 2019)	ACM MM 2019	VGG−M	ImageNet	88.2	70.7	–	–	76.6	53.7	43.1	38.3	31.4
MANet (Li et al. 2019b)	ICCVW 2019	VGG−M	ImageNet	89.4	72.4	–	–	77.7	53.9	45.5	–	32.6
CMPP (Wang et al. 2020)	CVPR 2020	VGG−M	ImageNet	92.6	73.8	–	–	82.3	57.5	–	–	–
CAT (Li et al. 2020)	ECCV 2020	VGG−M	ImageNet	88.9	71.7	79.2	53.3	80.4	56.1	45.0	39.5	31.4
APFNet (Xiao et al. 2022)	AAAI 2022	VGG−M	ImageNet	90.5	73.7	–	–	82.7	57.9	50.0	43.9	36.2
HMFT (Zhang et al. 2022)	CVPR 2022	ResNet−50	ImageNet	91.2	74.9	78.6	53.5	78.8	56.8	–	–	–
ProTrack (Yang et al. 2022)	ACM MM 2022	ViT−B	SOT	–	–	–	–	78.6	58.7	50.9	–	42.1
QAT (Liu et al. 2023a)	ACM MM 2023	ResNet−50	ImageNet	91.5	75.5	86.8	61.9	88.4	64.3	64.2	59.6	50.1
CMD (Zhang et al. 2023)	CVPR 2023	ResNet−50	ImageNet	89.2	73.4	–	–	82.4	58.4	59.0	54.6	46.4
ViPT (Zhu et al. 2023a)	CVPR 2023	ViT−B	SOT	–	–	–	–	83.5	61.7	65.1	–	52.5
TBSI (Hui et al. 2023)	CVPR 2023	ViT−B	SOT	–	–	85.3	62.5	87.1	63.7	69.2	65.7	55.6
STMT (Sun et al. 2024)	TCSVT 2024	ViT−B	SOT	–	–	83.0	59.5	86.5	63.8	67.4	63.4	53.7
TATrack (Wang et al. 2024)	AAAI 2024	ViT−B	SOT	–	–	85.3	61.8	87.2	64.4	70.2	66.7	56.1
BAT (Cao et al. 2024)	AAAI 2024	ViT−B	SOT	–	–	–	–	86.8	64.1	70.2	–	56.3
GMMT (Tang et al. 2024)	AAAI 2024	ViT−B	SOT	–	–	–	–	87.9	64.7	70.7	67.0	56.6
Un-Track (Wu et al. 2024)	CVPR 2024	ViT−B	SOT	–	–	–	–	83.7	61.8	66.7	–	53.6
SDSTrack (Hou et al. 2024)	CVPR 2024	ViT−B	SOT	–	–	–	–	84.8	62.5	66.5	–	53.1
OneTracker (Hong et al. 2024)	CVPR 2024	ViT−B	ImageNet	–	–	–	–	85.7	64.2	67.2	–	53.8
CKD (Lu et al. 2024)	ACM MM 2024	ViT−B	DropMAE	93.2	77.2	88.4	65.2	90.0	67.4	73.2	69.3	58.1
TPF (Lu et al. 2025a)	ARXIV 2025	ViT−B	DropMAE	94.3	76.3	88.0	63.8	89.7	67.1	75.1	71.3	59.5
AINet (Lu et al. 2025b)	AAAI 2025	ViT−B	DropMAE	–	–	86.8	64.1	89.1	66.8	73.0	69.0	58.2
STTrack (Hu et al. 2025a)	AAAI 2025	ViT−B	SOT	–	–	–	–	89.8	66.7	76.0	–	60.3
CAFormer (Xiao et al. 2025)	AAAI 2025	ViT−B	SOT	91.8	76.9	85.6	63.2	88.3	66.4	70.0	66.1	55.6
MambaVT (Lai et al. 2025)	TCSVT 2025	Vim−S	ImageNet	94.1	75.3	88.0	63.7	88.9	65.8	73.0	69.5	57.9
XTrack (Tan et al. 2025b)	ICCV 2025	ViT−B	DropMAE	–	–	–	–	87.4	64.9	69.1	–	55.7
CADTrack	Ours	ViT−B	SOT	95.3	77.8	88.7	62.9	90.9	65.6	75.8	71.9	60.2
CADTrack	Ours	ViT−B	DropMAE	**95.8**	**78.3**	**91.2**	**65.4**	**92.8**	**67.7**	**77.7**	**73.3**	**61.3**

Table 1: Performance on four RGBT tracking benchmarks. The best results are in bold.

where $\tau \in \{Z_0, Z_t\}$. $\mathcal{G}_m^\tau$ is a linear projection layer and v is the offset magnitude. Then, we utilize bilinear sampling $\mathcal{U}$ to amplify discriminative features (Xia et al. 2022):

$$\mathbf{F}_m^\tau = \mathcal{U}\left(\hat{\mathbf{F}}_m^\tau, \tilde{P} + \Delta\tilde{P}_m^\tau\right). \tag{15}$$

Afterwards, we concatenate the sampled features to obtain $\mathbf{F}_S = \left[\mathbf{F}_m^{Z_0}; \mathbf{F}_m^{Z_t}\right]$. For temporal propagation, we update alignment cues via cross-modal cross-attention Φ:

$$\mathbf{C}_m^{t+1} = \mathbf{C}_m^t + \Phi\left(\mathbf{C}_m^t, \mathbf{F}_S\right). \tag{16}$$

Finally, we introduce the intra-modal cross-attention to refine features of search regions by three sequential steps: spatial guidance, feature enhancement and response generation. The corresponding formulations are as follows:

$$\hat{\mathbf{C}}_m^{t+1} = \mathbf{C}_m^{t+1} + \Phi(\mathbf{C}_m^{t+1}, \overline{\mathbf{F}}_m^{S_t}), \tag{17}$$

$$\tilde{\mathbf{C}}_m^{t+1} = \hat{\mathbf{C}}_m^{t+1} + \mathrm{FFN}(\hat{\mathbf{C}}_m^{t+1}), \tag{18}$$

$$\mathbf{H}_m^t = \overline{\mathbf{F}}_m^{S_t} \otimes (\tilde{\mathbf{C}}_m^{t+1})^T \odot \overline{\mathbf{F}}_m^{S_t}, \tag{19}$$

where $\otimes$ denotes the matrix multiplication.

Unlike previous methods that rely on spatial alignment, our DAM generates spatiotemporal alignment cues through deformable sampling and temporal propagation. It preserves modality-specific spatial characteristics while enhancing complementary information. In addition, by leveraging modality-specific offsets and cross-modal cross-attention, our DAM addresses spatial misalignment caused by modality discrepancies without explicit supervision.

Experiments

Implementation Details

We implement CADTrack based on the PyTorch toolbox and train it on 4 NVIDIA V100 GPUs, with a global batch size of 32. The framework adopts a ViT-B backbone initialized from SOT (Ye et al. 2022) and DropTrack (Wu et al. 2023). Training employs the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate 10^{-4} and a weight decay 10^{-4}. Our MFI is deployed at the 4-th, 7-th, and 10-th layers of ViT-B, employing channel compression with 8. We select 6 experts per modality. $N_K = 1$ and $v = 5$ to balance deformation adaptability and localization precision. Template images and search regions utilize fixed resolutions of 128×128 and 256×256, respectively. For training, we use the training set of LasHeR (Li et al. 2021). On VTUAV (Zhang et al. 2022), we instead train exclusively on its training set.

Comparison with State-of-the-Art Trackers

We evaluate CADTrack on GTOT (Li et al. 2016), RGBT210 (Li et al. 2017), RGBT234 (Li et al. 2019a), LasHeR and VTUAV benchmarks using Precision Rate (PR) and Success Rate (SR), with additional Normalized Precision Rate (NPR) for LasHeR. Notably, ground truth annotations are misaligned in GTOT, RGBT234 and VTUAV. Following previous evaluation practices (Hu et al. 2025b; Shao et al. 2025) , we employ Maximum Precision Rate (MPR) and Maximum Success Rate (MSR) on these benchmarks.

GTOT. As shown in Tab. 1, our method shows outstanding performance with 95.8% MPR and 78.3% MSR. It outperforms CAFormer (Xiao et al. 2025) by a margin of +4.0% MPR and shows improved robustness than the early-fusion

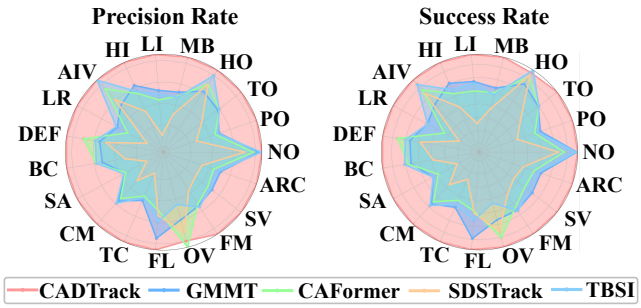

Figure 6: Attribute-based evaluation on the LasHeR dataset.

Method	VTUAV-ST		VTUAV-LT	
	MPR ↑	MSR ↑	MPR ↑	MSR ↑
DAFNet (Gao et al. 2019)	62.0	45.8	25.3	18.8
mfDiMP (Zhang et al. 2019)	67.3	55.4	31.5	27.2
HMFT (Zhang et al. 2022)	75.8	62.7	41.4	35.5
HMFT_LT (Zhang et al. 2022)	-	-	53.6	46.1
QAT (Liu et al. 2023a)	80.1	66.7	-	-
CKD (Lu et al. 2024)	90.2	77.8	-	-
AINet (Lu et al. 2025b)	87.1	74.5	-	-
CAFormer (Xiao et al. 2025)	88.6	76.2	-	-
MambaVT (Lai et al. 2025)	88.6	75.7	-	-
CADTrack(Ours)	**90.4**	**78.2**	**61.3**	**53.7**

Table 2: Performance on VTUAV benchmark.

method TPF (Lu et al. 2025a) with a margin of +2.0% MSR.

RGBT210. As shown in Tab. 1, our method achieves 91.2% PR and 65.4% SR. Specifically, it outperforms TATrack (Wang et al. 2024) by 5.9% in PR through cross-modal feature interaction, while surpassing STMT (Sun et al. 2024) by 5.9% in SR, validating the effectiveness of our method in handling modality-specific challenges.

RGBT234. As shown in Tab. 1, our method achieves leading performance, surpassing recent trackers: +5.4% MPR over XTrack (Tan et al. 2025b) and +5.2% MSR over SDSTrack (Hou et al. 2024). These results demonstrate that our method has an outstanding cross-modal feature representation capability, particularly in handling challenging scenarios with modality discrepancies.

LasHeR. As shown in Tab. 1, our method delivers state-of-the-art performance with PR 77.7% and SR 61.3%. These results demonstrate notable improvements: +7.0% PR over GMMT (Tang et al. 2024), and +7.7% SR over UnTrack (Wu et al. 2024). Fig. 6 further reveals the attribute-specific dominance, with significant improvements in motion blur (MB) and deformation (DEF). Notably, substantial gains in thermal crossover (TC) and low illumination (LI) highlight the robustness under challenging conditions.

VTUAV. As shown in Tab. 2, on the short-term (ST) subset, our method achieves 78.2% MSR, outperforming AINet (Lu et al. 2025b) by +3.7%. On the long-term (LT) subset, it attains 61.3% MPR, surpassing HMFT_LT (Zhang et al. 2022) by +7.7%. These results highlight a balanced accuracy-efficiency across both ST and LT scenarios.

Ablation Studies

In this subsection, we conduct ablation experiments to assess the effect of different components.

Method	Pretraining	PR↑	NPR↑	SR↑
Baseline	SOT	67.9	64.4	54.5
+ Template update	SOT	69.1	65.6	55.6
+ DAM	SOT	72.7	69.3	58.3
+ MFI	SOT	74.1	70.4	59.2
+ CAM	SOT	75.8	71.9	60.2
Full model	DropMAE	**77.7**	**73.3**	**61.3**

Table 3: Ablation study of key components on LasHeR.

Module	PR↑	NPR↑	SR↑	Params↓	FLOPs↓	FPS↑
TBSI	74.5	70.9	59.7	229.2M	108.8G	29
BAT	75.9	72.1	60.4	130.1M	77.4G	39
BSI	76.9	72.6	60.8	133.1M	79.8G	35
MFI	**77.7**	**73.3**	**61.3**	**130.0M**	**77.3G**	**40**

Table 4: Comparison of interaction mechanisms on LasHeR.

4	7	10	PR↑	NPR↑	SR↑
✓			75.2	71.4	59.8
✓	✓		75.9	71.7	60.3
✓	✓	✓	**77.7**	**73.3**	**61.3**

Table 5: Impact of interaction positions on LasHeR.

Component Analysis. As shown in Tab. 3, the baseline (ViT-B+convolutional fusion) achieves 67.9% PR, revealing fundamental limitations. Template update mechanisms address temporal drift by refreshing target representations, yielding a +1.2% PR. Our DAM resolves spatial misalignment through deformable sampling, delivering the gain of +3.6% PR. Our MFI enables an efficient modality interaction, resulting in +1.4% PR. Finally, our CAM dynamically activates backbone layers through sparse gating, achieving 75.8% PR. The use of pre-trained DropTrack provides additional gains, showing the generalization of our method.

Interaction Mechanism Analysis. We further validate the efficacy of MFI against other paradigms: (1) template-bridged search region interaction (TBSI) (Hui et al. 2023); (2) bi-directional adapter (BAT) (Cao et al. 2024); (3) background suppression interactive (BSI) (Hu et al. 2025a). As shown in Tab. 4, our MFI achieves an efficiency-performance balance, outperforming TBSI by +3.2% PR with 29% lower FLOPs and 38% higher FPS. This advantage stems from our linear-complexity feature interaction.

Effect of Interaction Positions. As shown in Tab. 5, the interactions of Layers 4, 7, and 10 achieves the best performance with 77.7% PR, demonstrating that early interactions capture spatial patterns while deeper interactions incorporate semantic abstractions. Our cross-layer design yields complementary advantages.

Effect of Mamba Layers in MFI. As shown in Tab. 6, the 2-layer configuration achieves optimal results with 77.7% PR. In contrast, the 1-layer variant shows a significant drop in PR by 2.3%. The 3-layer variant exhibits a minor decline in PR by 0.6%. This slight degradation suggests that the excessive model complexity may lead to over-fitting.

Effect of State Dimensions in MFI. The state dimension in SSM plays a key role in modeling long-range de-

Layer	PR↑	NPR↑	SR↑
1	75.4	71.4	60.1
2	**77.7**	**73.3**	**61.3**
3	77.1	72.6	61.0

Table 6: Effect of Mamba layers on LasHeR.

Selection Strategy	PR↑	NPR↑	SR↑
Top-k ($k = 4$)	75.5	71.4	59.9
Top-k ($k = 6$)	**77.7**	**73.3**	**61.3**
Top-k ($k = 8$)	74.3	70.2	58.9
Fixed interval	75.8	71.6	59.9
Manual selection	75.3	71.2	59.6

Table 7: Expert selection strategies of CAM on LasHeR.

N_K	PR↑	NPR↑	SR↑	ΔFLOPs
1	**77.7**	**73.3**	**61.3**	–
2	75.6	71.3	59.9	+0.3%
4	76.9	72.6	60.8	+0.8%

Table 8: Impact of cue quantity on LasHeR.

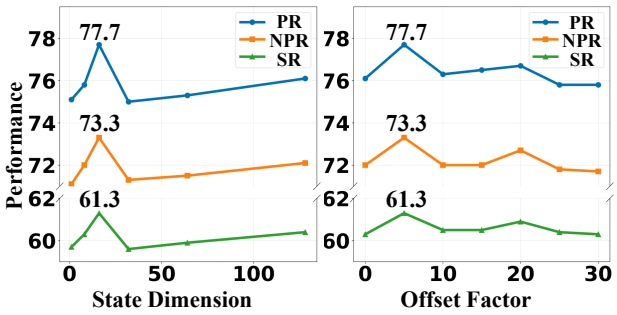

Figure 7: Comparison with different hyper-parameters.

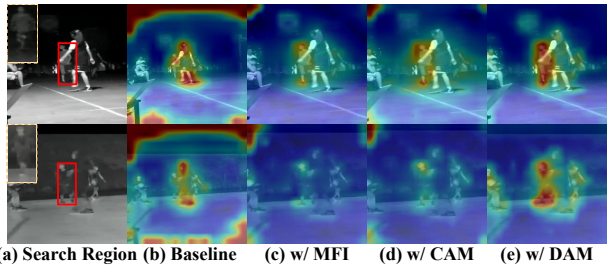

Figure 8: Attention evolution for RGB (top) and T (bottom).

pendencies. As shown in Fig. 7, the 16-dimensional configuration achieves the best performance with 77.7% PR. Smaller dimensions exhibit a reduced representational ability due to insufficient context encoding. Larger dimensions perform worse despite more parameters. These findings indicate that large state dimensions introduce redundancy, while insufficient dimensions limit context modeling. The 16-dimensional state balances expressiveness and computational efficiency, capturing essential spatiotemporal dependencies for cross-modal feature interaction.

Effect of Expert Numbers in CAM. As shown in Tab. 7, our method achieves optimal performance of 77.7% PR by selecting 6 experts. It achieves higher performance than the one with $k = 4$, while avoiding the increased noise observed at $k = 8$ which yields 74.3% PR. Compared methods show clear limitations: fixed-interval selection (every 2 layers) yields 75.8% PR due to the lack of adaptability in dynamic scenes, while the manual selection (1,3,6,9,12) obtains 75.3% PR due to the suboptimal spatial-semantic balance. Our method enables dynamic prioritization: shallow layers are activated for deformation scenarios, while deep layers focus on occlusion robustness. By leveraging this layer-specific activation, the model addresses diverse tracking challenges with improved efficiency and accuracy.

Effect of Cue Quantity in DAM. As evidenced in Tab. 8, the quantity of spatiotemporal alignment cues influences tracking precision. The optimal performance is achieved at $N_K = 1$ with 77.7% PR, significantly outperforming multi-cue configurations. The performance difference arises from two key factors: (1) excessive cues introduce conflicting signals due to over-parameterization, and (2) single-cue processing relies on temporal consistency to simplify state representations. Additionally, the configuration of $N_K = 1$ reduces the computational overhead of multi-cue variants.

Effect of the Offset Factor in DAM. The offset factor v determines the deformation range. As shown in Fig. 7, $v = 5$ achieves the best performance with 77.7% PR. Smaller values reduce the representation ability due to limited deformation ranges. In contrast, larger values introduce instability during rapid motion. These results highlight that large deformation ranges create inconsistent movement, while insufficient ranges limit adaptability. The configuration of $v = 5$ maintains tracking stability under motion variations.

Visualization Analysis

Fig. 8 illustrates the evolution of attention maps. The baseline model exhibits inconsistent attention, leading to wrong target localization. Our MFI aligns cross-modal features, enabling correct target localization. Our CAM focuses on key features, while DAM eliminates noise for robust tracking.

Conclusion

In this work, we propose a novel feature learning framework named CADTrack for RGBT tracking. For modality interaction, we introduce Mamba-based Feature Interaction (MFI), which achieves linear-complexity feature interaction while preserving contextual coherence. To enhance feature robustness, we propose the Contextual Aggregation Module (CAM), dynamically activating backbone layers through sparse gating and leveraging complementary information from cross-layer features. Furthermore, we propose the Deformable Alignment Module (DAM), resolving modality-specific misalignment via deformable sampling and temporal propagation. Extensive experiments on five benchmarks validate the effectiveness of our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62576069), Dalian Science and Technology Innovation Fund (No.2023JJ11CG001) and Natural Science Foundation of Liaoning Province (No.2025-MS-025).

References

Cai, W.; Liu, Q.; and Wang, Y. 2025. SPMTrack: Spatio-Temporal Parameter-Efficient Fine-Tuning with Mixture of Experts for Scalable Visual Tracking. In *CVPR*, 16871–16881.

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional Adapter for Multimodal Tracking. In *AAAI*, volume 38, 927–935.

Chen, Y.; and Wang, L. 2025. eMoE-Tracker: Environmental MoE-Based Transformer for Robust Event-Guided Object Tracking. *IEEE RAL*, 10(2): 1393–1400.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *NN*, 107: 3–11.

Fan, H.; Yu, Z.; Wang, Q.; Fan, B.; and Tang, Y. 2024. QueryTrack: Joint-modality Query Fusion Network for RGBT Tracking. *IEEE TIP*, 33: 3187–3199.

Feng, X.; Hu, S.; Li, X.; Zhang, D.; Wu, M.; Zhang, J.; Chen, X.; and Huang, K. 2025a. ATCTrack: Aligning Target-Context Cues with Dynamic Target States for Robust Vision-Language Tracking. In *ICCV*, 19850–19861.

Feng, X.; Zhang, D.; Hu, S.; Li, X.; Wu, M.; Zhang, J.; Chen, X.; and Huang, K. 2025b. CSTrack: Enhancing RGB-X Tracking via Compact Spatiotemporal Features. In *ICML*.

Gao, Y.; Li, C.; Zhu, Y.; Tang, J.; He, T.; and Wang, F. 2019. Deep Adaptive Fusion Network for High Performance RGBT Tracking. In *ICCVW*, 91–99.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining Recurrent, Convolutional, and Continuous-Time Models with Linear State Space Layers. In *NeurIPS*, volume 34, 572–585.

Guo, M.; Tan, W.; Ran, W.; Jing, L.; and Zhang, Z. 2025. DreamTrack: Dreaming the Future for Multimodal Visual Object Tracking. In *CVPR*, 7201–7210.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning. In *CVPR*, 19079–19091.

Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. 2024. SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-modal Visual Object Tracking. In *CVPR*, 26551–26561.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025a. Exploiting Multimodal Spatial-Temporal Patterns for Video Object Tracking. In *AAAI*, volume 39, 3581–3589.

Hu, X.; Zhong, B.; Liang, Q.; Shi, L.; Mo, Z.; Tai, Y.; and Yang, J. 2025b. Adaptive perception for unified visual multimodal object tracking. *IEEE TAI*, 6(10): 2819–2829.

Hu, Y.; Shao, Z.; Fan, B.; and Liu, H. 2025c. Dual-Level Modality De-Biasing for RGBT Tracking. *IEEE TIP*, 34: 2667–2679.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging Search Region Interaction With Template for RGBT Tracking. In *CVPR*, 13630–13639.

Lai, S.; Liu, C.; Zhu, J.; Kang, B.; Liu, Y.; Wang, D.; and Lu, H. 2025. MambaVT: Spatio-temporal Contextual Modeling for Robust RGBT Tracking. *IEEE TCSVT*, 35(9): 9312–9323.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking. *IEEE TIP*, 25(12): 5743–5756.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019a. RGBT Object Tracking: Benchmark and Baseline. *PR*, 96: 106977.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware RGBT Tracking. In *ECCV*, 222–237.

Li, C.; Lu, A.; Zheng, A.; Tu, Z.; and Tang, J. 2019b. Multi-adapter RGBT Tracking. In *ICCVW*, 2262–2270.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A Large-scale High-diversity Benchmark for RGBT Tracking. *IEEE TIP*, 31: 392–404.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted Sparse Representation Regularized Graph Learning for RGBT Object Tracking. In *ACM MM*, 1856–1864.

Li, N.; Zhong, B.; Liang, Q.; Mo, Z.; Nong, J.; and Song, S. 2025. SIEVL-Track: Exploring Semantic Information Enhancement for Visual-Language Object Tracking. *IEEE TCSVT*, 35(6): 5872–5884.

Liu, L.; Li, C.; Xiao, Y.; and Tang, J. 2023a. Quality-aware RGBT Tracking via Supervised Reliability Learning and Weighted Residual Guidance. In *ACM MM*, 3129–3137.

Liu, X.; Zhou, L.; Zhou, Z.; Chen, J.; and He, Z. 2025. Mambavlt: Time-evolving multimodal state space model for vision-language tracking. In *CVPR*, 8731–8741.

Liu, Y.; Zhang, S.; Chen, J.; Yu, Z.; Chen, K.; and Lin, D. 2023b. Improving Pixel-based MIM by Reducing Wasted Modeling Capability. In *ICCV*, 5361–5372.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, A.; Guo, Y.; Wang, W.; Li, C.; Tang, J.; and Luo, B. 2025a. Breaking Shallow Limits: Task-Driven Pixel Fusion for Gap-free RGBT Tracking. *arXiv preprint arXiv:2503.11247*.

Lu, A.; Wang, W.; Li, C.; Tang, J.; and Luo, B. 2025b. RGBT Tracking via All-layer Multimodal Interactions with Progressive Fusion Mamba. In *AAAI*, volume 39, 5793–5801.

Lu, A.; Zhao, J.; Li, C.; Xiao, Y.; and Luo, B. 2024. Breaking Modality Gap in RGBT Tracking: Coupled Knowledge Distillation. In *ACM MM*, 9291–9300.

Shao, Z.; Hu, Y.; Fan, B.; and Liu, H. 2025. PURA: Parameter Update-Recovery Test-Time Adaption for RGBT Tracking. In *CVPR*, 22089–22098.

Sun, D.; Pan, Y.; Lu, A.; Li, C.; and Luo, B. 2024. Transformer RGBT Tracking with Spatio-temporal Multimodal Tokens. *IEEE TCSVT*, 34(11): 12059–12072.

Tan, Y.; Shao, J.; Zamfir, E.; Li, R.; An, Z.; Ma, C.; Paudel, D. P.; Gool, L. V.; Timofte, R.; and Wu, Z. 2025a. What You Have is What You Track: Adaptive and Robust Multimodal Tracking. In *ICCV*, 3455–3465.

Tan, Y.; Wu, Z.; Fu, Y.; Zhou, Z.; Sun, G.; Zamfi, E.; Ma, C.; Paudel, D. P.; Gool, L. V.; and Timofte, R. 2025b. XTrack: Multimodal Training Boosts RGB-X Video Object Trackers. In *ICCV*, 5734–5744.

Tang, Z.; Xu, T.; Wu, X.; Zhu, X.-F.; and Kittler, J. 2024. Generative-based Fusion Mechanism for Multi-modal Tracking. In *AAAI*, volume 38, 5189–5197.

Tang, Z.; Xu, T.; Wu, X.-J.; Zhu, X.; Cheng, C.; Feng, Z.; and Kittler, J. 2025. Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution. *IEEE TIP*, 34: 7235–7249.

Wang, C.; Xu, C.; Cui, Z.; Zhou, L.; Zhang, T.; Zhang, X.; and Yang, J. 2020. Cross-modal Pattern-propagation for RGB-T Tracking. In *CVPR*, 7064–7073.

Wang, H.; Liu, X.; Li, Y.; Sun, M.; Yuan, D.; and Liu, J. 2024. Temporal Adaptive RGBT Tracking with Modality Prompt. In *AAAI*, volume 38, 5436–5444.

Wu, Q.; Yang, T.; Liu, Z.; Wu, B.; Shan, Y.; and Chan, A. B. 2023. DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks. In *CVPR*, 14561–14571.

Wu, Z.; Zheng, J.; Ren, X.; Vasluianu, F.-A.; Ma, C.; Paudel, D. P.; Gool, L. V.; and Timofte, R. 2024. Single-model and Any-modality for Video Object Tracking. In *CVPR*, 19156–19166.

Xia, J.; Shi, D.; Song, K.; Song, L.; Wang, X.; Jin, S.; Zhao, C.; Cheng, Y.; Jin, L.; Zhu, Z.; Li, J.; Wang, G.; Xing, J.; and Zhao, J. 2024. Unified Single-Stage Transformer Network for Efficient RGB-T Tracking. In *IJCAI*, 1471–1479.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision Transformer with Deformable Attention. In *CVPR*, 4794–4803.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based Progressive Fusion Network for RGBT Tracking. In *AAAI*, volume 36, 2831–2838.

Xiao, Y.; Zhao, J.; Lu, A.; Li, C.; Yin, B.; Lin, Y.; and Liu, C. 2025. Cross-modulated Attention Transformer for RGBT Tracking. In *AAAI*, volume 39, 8682–8690.

Xu, T.; Kang, Z.; Zhu, X.; and Wu, X.-J. 2024. Learning adaptive spatio-temporal inference transformer for coarse-to-fine animal visual tracking: algorithm and benchmark. *IJCV*, 132(7): 2698–2712.

Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Prompting for Multi-modal Tracking. In *ACM MM*, 3492–3500.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint Feature Learning and Relation Modeling for Tracking: A One-stream Framework. In *ECCV*, 341–357.

Yuksel, S. E.; Wilson, J. N.; and Gader, P. D. 2012. Twenty years of mixture of experts. *IEEE TNNLS*, 23(8): 1177–1193.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; Weijer, J. V. D.; and Khan, F. S. 2019. Multi-modal Fusion for End-to-end RGB-T Tracking. In *ICCVW*, 2252–2261.

Zhang, P.; Zhao, J.; Bo, C.; Wang, D.; Lu, H.; and Yang, X. 2021. Jointly Modeling Motion and Appearance Cues for Robust RGB-T Tracking. *IEEE TIP*, 30: 3335–3347.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-Thermal UAV Tracking: A Large-scale Benchmark and New Baseline. In *CVPR*, 8886–8895.

Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; and Han, J. 2023. Efficient RGB-T Tracking via Cross-modality Distillation. In *CVPR*, 5404–5413.

Zhang, T.; He, X.; Jiao, Q.; Zhang, Q.; and Han, J. 2024. AMNet: Learning to Align Multi-modality for RGB-T Tracking. *IEEE TCSVT*, 34(8): 7386–7400.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023a. Visual Prompt Multi-modal Tracking. In *CVPR*, 9516–9526.

Zhu, X.-F.; Xu, T.; Atito, S.; Awais, M.; Wu, X.-J.; Feng, Z.; and Kittler, J. 2024a. Self-supervised learning for RGB-D object tracking. *PR*, 155: 110543.

Zhu, X.-F.; Xu, T.; Liu, Z.; Tang, Z.; Wu, X.-J.; and Kittler, J. 2024b. UniMod1K: Towards a more universal large-scale dataset and benchmark for multi-modal learning. *IJCV*, 132(8): 2845–2860.

Zhu, X.-F.; Xu, T.; Tang, Z.; Wu, Z.; Liu, H.; Yang, X.; Wu, X.-J.; and Kittler, J. 2023b. RGBD1K: A Large-scale Dataset and Benchmark for RGB-D Object Tracking. In *AAAI*, volume 37, 3870–3878.

Zhu, Y.; Li, C.; Luo, B.; Tang, J.; and Wang, X. 2019. Dense Feature Aggregation and Pruning for RGBT Tracking. In *ACM MM*, 465–472.

Zhu, Z.; Zhong, B.; Liang, Q.; Yang, H.; Zheng, Y.; and Li, N. 2025. Adaptive Expert Decision for RGB-T Tracking. *IEEE TCSVT*, 35(10): 10330–10338.