

Local-to-global Cross-modal Coordination for Self-supervised RGB-T Tracking

Yueying Zhang^{1*}, Timing Li^{1*}, Bing Cao^{2†}, and Pengfei Zhu^{2,3,4}

¹ School of Computer Science and Technology, Tianjin University, Tianjin, China

² School of Artificial Intelligence, Tianjin University, Tianjin, China

³ Low-Altitude Intelligence Laboratory, Xiong'an National Innovation Center, Hebei, China

⁴ Xiong'an Guochuang Lantian Technology Co., Ltd., Hebei, China
{zhangyueying, litm, caobing, zhupengfei}@tju.edu.cn

Abstract. Self-supervised RGB-T tracking seeks to learn robust cross-modal representations without dense manual annotations. A common supervision strategy relies on forward-backward tracking loops, but in practice, it is highly vulnerable to error accumulation and modality drift caused by the inconsistent reliability of visible and thermal cues. To address this issue, we propose LGCTrack, a self-supervised RGB-T tracker underpinned by a local-to-global coordination strategy within a closed-loop verification paradigm. At the local level, a hierarchical coordination module performs layer-wise, reliability-aware token interaction to enable adaptive cross-modal fusion. At the global level, a turning-phase coordination module refines the target representation during the critical forward-to-backward transition, anchoring target semantics and improving backward tracking. By jointly coordinating cross-modal interaction at two granularities, LGCTrack intrinsically filters pseudo-label noise and mitigates drift. Experiments on standard RGB-T tracking benchmarks show that LGCTrack consistently outperforms existing self-supervised methods and substantially narrows the gap to fully supervised trackers. Our code is available at <https://github.com/zyyyya/LGCTrack>.

Keywords: RGB-T tracking · Single object tracking · Self-supervised learning

1 Introduction

Single-object tracking is a fundamental task in computer vision, with applications in video surveillance, robotics, and autonomous driving [3, 6, 13]. Recent RGB trackers have achieved substantial progress with the development of transformer-based architectures [23, 46, 48, 55] and large-scale datasets [5, 11, 30]. Nevertheless, their performance remains limited under adverse imaging conditions, such as low illumination, heavy shadow, and severe appearance variation, where the visible modality alone often becomes insufficient for accurate target

*Equal contribution. †Corresponding author.

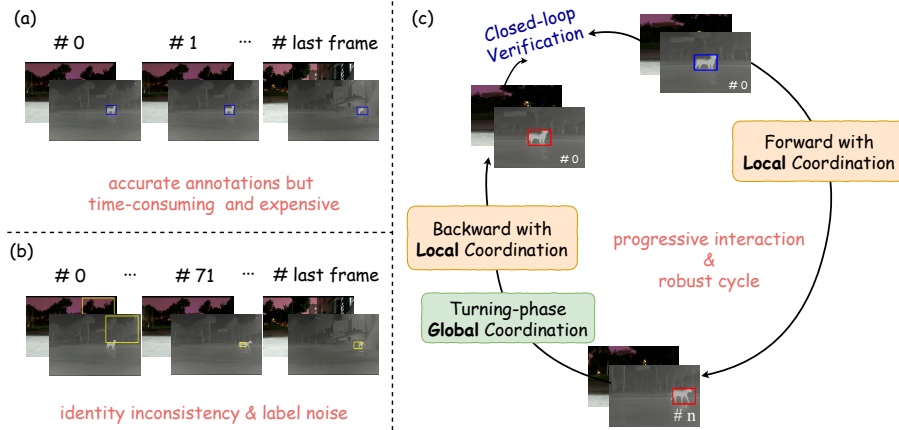


Fig. 1: (a) Fully supervised tracker utilize accurate annotations but manual annotations are time-consuming and expensive. (b) Pseudo-labels often suffer from identity inconsistency and label noise. (c) We achieve local and global cross-modal coordination under closed-loop verification. Blue, yellow, and red boxes indicate the ground-truth label, pseudo-label, and predicted box, respectively.

localization [36]. RGB-T tracking addresses this limitation by incorporating thermal infrared image as an additional modality. Since thermal data is less sensitive to illumination changes, it provides complementary information when RGB observations are degraded, making RGB-T tracking more reliable in challenging scenarios [1, 56].

Although fully supervised RGB-T trackers have achieved impressive performance [2, 10, 29, 34, 35, 45, 47], most of them rely on dense frame-level annotations for training, as illustrated in Fig. 1 (a). Constructing large scale RGB-T datasets with accurate manual labels is time-consuming and expensive, which substantially limits the scalability of supervised learning. This limitation has recently stimulated growing interest in self-supervised RGB-T tracking [20, 21], where pseudo-labels are typically generated from the temporal continuity of videos.

However, due to the dynamic discrepancies between modalities [7, 24], these pseudo-labels are often unreliable, as illustrated in Fig. 1 (b). As a result, cross-modal interaction becomes inherently challenging. While it is essential for exploiting complementary information across modalities, inaccurate cues from one modality may also propagate to the other and lead to error accumulation. This issue is particularly pronounced in closed-loop verification based frameworks, where the target is first tracked from a reference frame to a subsequent frame and then back to the original frame. Unreliable modal information is introduced during this process, the resulting errors can not only disrupt forward tracking but also corrupt the target representation, thereby weakening the closed-loop verification [26, 29].

The central difficulty of loop-based self-supervised RGB-T tracking lies in how cross-modal interaction is regulated throughout the tracking loop. This difficulty arises at two closely connected levels. During feature extraction, cross-modal information exchange should be sufficiently effective to exploit modality complementarity, while remaining controllable enough to avoid transferring unreliable cues. At the transition from forward tracking to backward tracking, the target representation must be further stabilized, since its quality directly affects the subsequent backward tracking. This calls for a coordinated solution that addresses both progressive feature interaction and target representation refinement within a unified framework.

To address this challenge, we propose **LGCTrack**, a **Local-to-Global Cross-modal** coordination framework for self-supervised RGB-T tracking, as shown in Fig. 1 (c). LGCTrack regulates cross-modal interaction during both feature extraction and the transition from forward tracking to backward tracking. Specifically, a hierarchical local coordination module is introduced into the dual-stream backbone to support progressive cross-modal exchange while limiting the transfer of unreliable cues. In addition, a turning-phase global coordination module is designed to refine the target representation before backward tracking, thereby improving the stability of the backward trajectory. Through this local-to-global coordination strategy, LGCTrack reduces error accumulation in the tracking loop and enhances the robustness of self-supervised RGB-T tracking.

The main contributions of this work are summarized as follows:

- We propose LGCTrack, a self-supervised RGB-T tracking framework anchored in a novel closed-loop verification mechanism. By effectively circumventing the unreliability of noisy pseudo-labels, it facilitates accurate spatio-temporal self-alignment of cross-modal features.
- We introduce a dual-stage coordination mechanism, comprising Hierarchical Local Coordination (HLC) for granular feature fusion and Turning-phase Global Coordination (TGC) to neutralize modal biases, ensuring stable and discriminative target representations.
- LGCTrack consistently outperforms existing self-supervised methods on multiple benchmarks and rivals fully-supervised counterparts, demonstrating exceptional robustness against temporal error accumulation.

2 Related Work

Self-supervised tracking. Self-supervised tracking learns target representations from unlabeled data without relying on dense manual annotations [22, 33, 49]. Most existing progress has been made in RGB tracking, where supervision is typically derived from temporal continuity, cycle-consistency, or discriminative learning objectives [40]. Early works such as UDT [39] and USOT [53] formulate tracking through forward-backward consistency. PUL [41] further introduces progressive discriminative learning with temporal correspondence mining and contrastive background discrimination. More recently, SSTrack [54] combines cycle-based association with instance-level contrastive supervision, while

Diff-Tracker [52] explores diffusion models as a source of localization cues. These methods show that cycle-based learning provides an effective foundation for self-supervised tracking, but they are developed for single-modality RGB tracking and do not consider cross-modal interaction under dynamic modality reliability.

RGB-T tracking. RGB-T tracking improves robustness under adverse imaging conditions by combining visible and thermal modalities [15, 43, 50]. Recent fully supervised methods achieve strong performance through increasingly effective cross-modal modeling [29, 32]. ViPT [56] adapts pre-trained RGB models to multimodal tracking via visual prompt learning. BAT [1] introduces a bidirectional adapter for dynamic modality adaptation. AINet [29] performs progressive all-layer fusion, while STTrack [10] models multimodal spatio-temporal correlations with a temporal state generator. Despite their effectiveness, these methods rely on dense supervision to correct inaccurate cross-modal fusion. Recent studies such as GDSTrack [20] have started to explore self-supervised RGB-T tracking. In self-supervised RGB-T tracking, however, unreliable cues may be transferred across modalities and further accumulate throughout the forward-backward loop. The regulation of cross-modal interaction throughout the overall tracking loop remains insufficiently studied. Different from prior methods that mainly pursue stronger fusion [19], our method focuses on controllable cross-modal coordination for loop-based self-supervised RGB-T tracking.

3 Method

In this section, we first define the task and give an overview of our framework (Sec. 3.1). Then we illustrate our local-to-global cross-modal coordination strategy, including hierarchical local coordination (Sec. 3.2) and turning-phase global coordination (Sec. 3.3).

3.1 Task Definition and Framework Overview

Task definition. Given a synchronized pair of visible and thermal video streams with an initial target specification, a self-supervised RGB-T tracker aims to accurately and consistently localize the same target in subsequent frames, learning to exploit multimodal information for tracking without relying on dense frame-level supervision during training.

Framework overview. The overall pipeline of our proposed self-supervised RGB-T tracking framework is shown in Fig. 2.

We formulate a foundational tracking operator, denoted as $\mathcal{T}$, which processes visible and thermal inputs in parallel to predict a target bounding box for each frame. For a given frame i , let $Z = (z^{rgb}, z^{tir})$ and $X_i = (x_i^{rgb}, x_i^{tir})$ denote the token embeddings of the template and search region in their respective modalities. Each modality stream maintains a global token denoted as

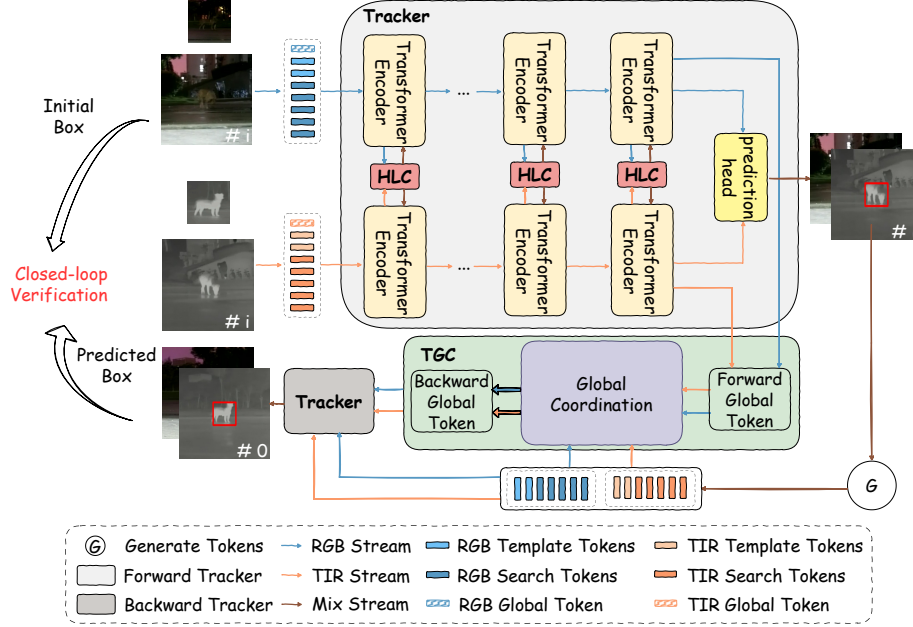


Fig. 2: The overall framework of LGCTrack. During forward tracking, a shared backbone extracts target features, where HLC interacts modality information at each encoder layer. At the turning phase, TGC coordinates cross-modal global tokens and initializes reverse tracking. The backward tracking then returns to the initial frame, providing a closed-loop verification signal for end-to-end optimization.

$S = (s^{rgb}, s^{tir})$, prepended to the image tokens. S encodes a compact representation of the current target state and is used for box prediction [55]. These tokens are processed by a dual-stream encoder $\phi(\cdot)$ with shared weights across modalities, maintaining modality-specific representations for RGB and TIR.

Specifically, we introduce our Hierarchical Local Coordination (HLC) inside the encoder (Sec. 3.2). At each layer, HLC enables controlled information exchange between the two streams. It leverages complementary cues while mitigating modality-dependent degradations. The resulting features are subsequently decoded by a prediction head $H(\cdot)$ to estimate the target state $\hat{b}_i$. Formally, this tracking process can be formulated as:

$$\hat{b}_i = \mathcal{T}(S, Z, X_i) = H(\phi(S, Z, X_i)). \quad (1)$$

Building upon this dual-stream tracker, we adopt a forward-backward tracking paradigm to learn tracking capabilities without dense annotations. Given an unannotated video clip of N frames and the initial target box b_{t_0} , the training proceeds through a closed loop.

In the forward tracking process, the forward tracker sequentially localizes the target. Initialized by the template region Z generated from b_{t_0} , the tracker

dynamically updates the search region X_i conditioned on the prediction from the preceding frame, $\hat{b}_{i-1}$. Using our defined operator $\mathcal{T}$, the forward trajectory is recursively generated as:

$$\hat{b}_i = \mathcal{T}(S, Z, X_i(\hat{b}_{i-1})), \quad i = t_0 + 1, t_0 + 2, \dots, N. \quad (2)$$

At the turning phase between forward and backward tracking, the backward tracking is initialized by the forward results from two aspects. Spatially, the backward template input $\tilde{Z}$ is initialized by the forward tracking estimate $\hat{b}_N$. Semantically, we apply a turning-phase global coordination(TGC) principle (Sec. 3.3). Instead of propagating the forward global tokens to the backward tracking in a direct way, we update the global tokens $\mathbf{s}_{turn}^{rgb}$ and $\mathbf{s}_{turn}^{tir}$ via our coordination module $\mathcal{Q}$:

$$\tilde{\mathbf{S}} = \tilde{\mathbf{s}}_{turn}^{rgb}, \tilde{\mathbf{s}}_{turn}^{tir} = \mathcal{Q}(\mathbf{s}_{turn}^{rgb}, \mathbf{s}_{turn}^{tir}). \quad (3)$$

The updated global tokens are then fed into the backward tracker together with the template $\tilde{Z}$. In the backward tracking process, we run the same tracker, the forward tracker $\mathcal{T}$, but in reverse temporal order. Conditioned on the new template $\tilde{Z}$ and the updated global token $\tilde{\mathbf{s}}$, the model iteratively predicts the backward result $\tilde{b}_{i-1}$:

$$\tilde{b}_{i-1} = \mathcal{T}(\tilde{\mathbf{S}}, \tilde{Z}, \tilde{X}_{i-1}(\tilde{b}_i)), \quad i = N, N-1, \dots, t_0 + 1, \quad (4)$$

where the initial backward state is set as $\tilde{b}_N = \hat{b}_N$, and $\tilde{X}_{i-1}(\tilde{b}_i)$ denotes the search region in frame $i-1$ cropped according to the backward prediction $\tilde{b}_i$ from frame i . The backward trajectory is then recursively generated from frame N to frame t_0 , with the updated global tokens providing consistent guidance throughout the loop.

By requiring the backward prediction to match the initial target location, the proposed framework establishes a closed-loop verification signal for optimization. The overall objective is defined as

$$\mathcal{L} = \mathcal{L}_{cls}(\tilde{b}_{t_0}, b_{t_0}^{gt}) + \lambda_{giou} \mathcal{L}_{giou}(\tilde{b}_{t_0}, b_{t_0}^{gt}) + \lambda_{L1} \mathcal{L}_1(\tilde{b}_{t_0}, b_{t_0}^{gt}), \quad (5)$$

where $\tilde{b}_{t_0}$ denotes the backward prediction at frame t_0 , and $b_{t_0}^{gt}$ is the corresponding ground-truth box. $\mathcal{L}_{cls}$ denotes the focal loss for classification [14]. $\mathcal{L}_{giou}$ and $\mathcal{L}_1$ denote the GIoU loss and L1 loss for box regression [31]. Following [48], λ_{giou} and λ_{L1} are set to 2 and 5, respectively.

3.2 Hierarchical Local Coordination

Our framework adopts a local-to-global coordination strategy. We regulate cross-modal interaction at two granularities: hierarchical local coordination inside the encoder (Sec. 3.2), and turning-phase global coordination at the forward-backward turning phase. (Sec. 3.3).

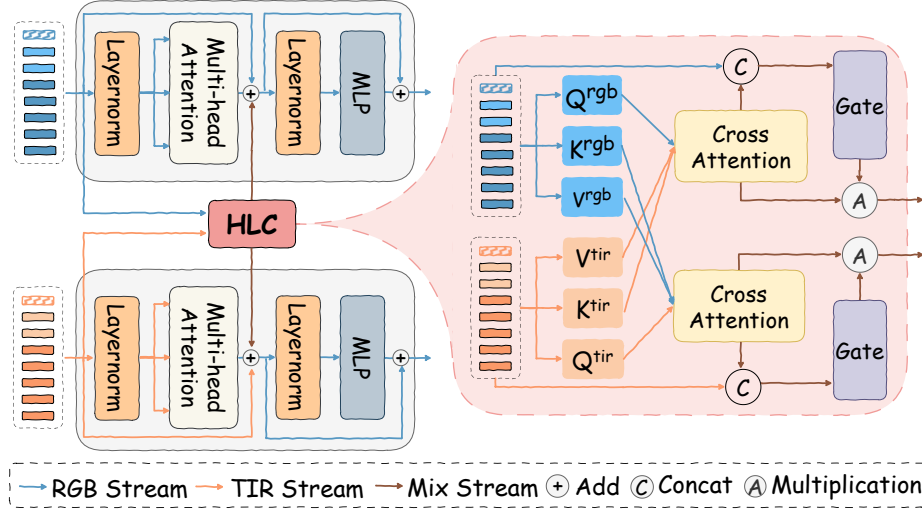


Fig. 3: The detailed framework of HLC.

At the local level, we introduce a Hierarchical Local Coordination (HLC) module within each block of the dual-stream encoder. As shown in Fig. 3, HLC is implemented as a bypass branch. For the l -th layer, let $F_{rgb}^{(l)}$ and $F_{tir}^{(l)}$ denote the intermediate token representations, HLC extracts cross-modal cues and injects them back to each stream with an explicit gate.

We first project query, key, and value projections for both modalities, denoted as $(Q_{rgb}, K_{rgb}, V_{rgb})$ and $(Q_{tir}, K_{tir}, V_{tir})$. To capture inter-modality dependencies, the queries of one modality attend to the keys and values of the other. For modality $m \in \{rgb, tir\}$ and its counterpart $\bar{m}$, the cross-modal feature is computed as

$$C_m = \mathcal{M} \left(\text{Softmax} \left(\frac{Q_m K_{\bar{m}}^T}{\sqrt{d}} \right) V_{\bar{m}} \right), \quad (6)$$

where d is the projection dimension, and $\mathcal{M}(\cdot)$ denotes a lightweight network implemented as a linear layer followed by a GELU activation layer.

In a self-supervised forward-backward loop, there is no per-frame supervision to correct intermediate interaction. When one modality becomes unreliable, directly adding cross-modal features can pass unreliable cues to the other stream. We therefore use an adaptive gate to control how much cross-modal information is injected. Specifically, we concatenate the original features with the cross-modal features and use a two-layer network (Linear-GELU-Linear) to output a single scalar gate value for each token.

$$g_m = \sigma \left(W_{gate}^m [F_m^{(l)}; C_m] \right), \quad m \in \{rgb, tir\}, \quad (7)$$

where $[\cdot; \cdot]$ denotes channel-wise concatenation, W_{gate}^m denotes the gating network parameters for modality m , and $\sigma(\cdot)$ is the sigmoid function.

The gated cross-modal features are then integrated back into the main processing stream as residuals with g_{rgb} and g_{tir} . We further introduce a learnable scaling parameter α_m , initialized to zero, to stabilize training. The updated feature for modality m is obtained by combining the spatial self-attention and the local cross-modal coordination:

$$F'_m = F_m^{(l)} + \text{MSA}\left(\text{LN}(F_m^{(l)})\right) + \alpha_m g_m \odot C_m, \quad m \in \{rgb, tir\}, \quad (8)$$

where $\text{LN}(\cdot)$ denotes Layer Normalization, MSA is the multi-head self-attention, and $\odot$ is element-wise multiplication. The intermediate feature is then updated by the Feed-Forward Network (FFN):

$$F_m^{(l+1)} = F'_m + \text{FFN}(\text{LN}(F'_m)), \quad m \in \{rgb, tir\}. \quad (9)$$

By applying this coordination process across all encoder layers, the encoder performs progressive local cue exchange with controlled information flow.

3.3 Turning-phase Global Coordination

While the HLC module can regulate feature exchange, closed-loop verification remains vulnerable at the turning phase from forward to backward tracking: the backward trajectory is initialized by the final result of the forward tracking, target drift at this switch can be propagated along the entire return path.

In a tracking framework with global tokens, the independent global tokens at the end of forward tracking are typically propagated to the next frame [55]. However, propagating these unimodal global tokens preserves modality-specific drift accumulated at the turning phase under the loop setting. To address this issue, we introduce a Turning-phase Global Coordination module (TGC).

Specifically, let $\mathbf{s}_{turn}^{rgb}$ and $\mathbf{s}_{turn}^{tir}$ denote the global tokens at the end of the forward tracking which aggregate target information [55], and let F_N^{rgb} , F_N^{tir} denote the corresponding image tokens of the initial frame of the backward tracking. Before reversing the tracking direction, we update global tokens via cross-modal attention so that it incorporates context from the other modality:

$$\tilde{\mathbf{s}}_{turn}^m = \text{Proj}\left(\text{Softmax}\left(\frac{Q(\mathbf{s}_{turn}^m)K(F_N^{\bar{m}})^T}{\sqrt{d}}\right)V(F_N^{\bar{m}})\right), m \in \{rgb, tir\}. \quad (10)$$

where $\bar{m}$ denotes the other modality, $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are modality-specific linear projections, and $\text{Proj}(\cdot)$ represents the final output linear projection. TGC updates the global tokens by aggregating spatial tokens from the other modality, and makes it explicitly incorporate cross-modal context at the turning phase.

To construct the cross-modal loop, we then use the updated global tokens to initialize backward tracking. $\tilde{\mathbf{s}}_{turn}^{rgb}$ is used on the TIR branch and $\tilde{\mathbf{s}}_{turn}^{tir}$ is used on the RGB branch. The cross-write initialization encourages the two streams to be coupled under the closed-loop verification objective, since the backward tracking of one modality depends on global tokens that are conditioned on the

other. As a result, the model is less likely to rely solely on unimodal states at the turning phase, and is encouraged to learn cross-modal representations that are useful for both directions. Combined with HLC, this design makes cross-modal interaction more controlled within the loop and improves the robustness of closed-loop verification. During inference, this mechanism simply serves to update the global token from the previous frame.

4 Experiments

4.1 Experiment Setting

Evaluation metrics. We evaluate LGCTrack on three RGB-T tracking benchmarks, including LasHeR [18], RGBT234 [17], and GTOT [16]. Following the standard evaluation protocols, we report Precision Rate (PR), Normalized Precision Rate (NPR), and Success Rate (SR) on LasHeR, and Precision Rate (PR) and Success Rate (SR) on both RGBT234 and GTOT. For the attribute-based evaluation on RGBT234, we further report the results under five challenge attributes, including low illumination (LI), low resolution (LR), no occlusion (NO), thermal crossover (TC), and background clutter (BC).

Implementation details. All quantitative results reported in this paper are obtained using NVIDIA RTX 4090 GPUs. Additionally, we reproduce the complete pipeline on Atlas 800T A3. The template and search inputs are resized to 128×128 and 256×256 , respectively. We adopt ViT-Base [4] as the feature extractor for both RGB and thermal streams and initialize our framework with SSTrack [54]. We adopt AdamW [27] optimizer with the learning rate of 10^{-4} and the weight decay of 10^{-4} . The model is trained for 10 epochs with a batch size of 4, and each epoch samples 3×10^4 training pairs.

4.2 Main Results

To verify its effectiveness, LGCTrack is comprehensively evaluated on LasHeR [18], RGBT234 [17], and GTOT [16], competing with contemporary RGB-T trackers across both fully supervised and self-supervised settings. As shown in Tab. 1, LGCTrack consistently delivers the best performance among self-supervised methods on all three benchmarks. Furthermore, it outperforms previous self-supervised approaches by clear margins across all reported metrics. LGCTrack effectively narrows the gap with supervised methods, achieving competitive or even superior performance without requiring manual annotations.

LasHeR. LasHeR [18] is a large-scale RGB-T tracking benchmark comprising 1,224 visible-thermal video pairs across 32 object categories, totaling 734.8K frames. As summarized in Tab. 1, LGCTrack attains 59.5% PR, 55.3% NPR, and 45.7% SR, establishing the best performance among all self-supervised methods.

Table 1: State-of-the-art comparison on LasHeR, RGBT234, and GTOT. “Sup” denotes whether the model is trained with Full supervision or Self-supervision.

Method	Source	Sup	LasHeR [18]			RGBT234 [17]		GTOT [16]	
			PR $\uparrow$	NPR $\uparrow$	SR $\uparrow$	PR $\uparrow$	SR $\uparrow$	PR $\uparrow$	SR $\uparrow$
APFNet [44]	AAAI 22	Full	50.0	-	36.2	82.7	57.9	90.5	73.7
CMD [51]	CVPR 23	Full	59.0	54.6	46.4	82.4	58.4	89.2	73.4
ViPT [56]	CVPR 23	Full	65.1	-	52.5	83.5	61.7	-	-
TBSI [12]	CVPR 23	Full	69.2	65.7	55.6	87.1	63.7	-	-
QAT [26]	ACM MM 23	Full	64.2	59.6	50.1	88.4	64.3	91.5	75.5
BAT [1]	AAAI 24	Full	70.2	-	56.3	86.8	64.1	-	-
TATrack [38]	AAAI 24	Full	70.2	66.7	56.1	87.2	64.4	-	-
GMMT [37]	AAAI 24	Full	70.7	67.0	56.6	87.9	64.7	93.6	78.5
SDSTrack [9]	CVPR 24	Full	66.5	-	53.1	84.8	62.5	-	-
OneTracker [8]	CVPR 24	Full	67.2	-	53.8	85.7	64.2	-	-
Un-Track [42]	CVPR 24	Full	66.7	-	53.6	84.2	62.5	-	-
CAT++ [25]	TIP 24	Full	50.9	44.4	35.6	84.0	59.2	91.5	73.3
IPL [28]	IJCV 24	Full	69.4	65.6	55.3	88.3	65.7	-	-
CAFormer [45]	AAAI 25	Full	70.0	66.1	55.6	88.3	66.4	91.8	76.9
SMSTracker [2]	ICCV 25	Full	70.3	-	56.0	86.9	64.5	-	-
XTrack [35]	ICCV 25	Full	69.1	-	55.7	87.4	64.9	-	-
S2OTFormer [21]	TOMM 24	Self	39.8	-	29.5	68.4	47.7	83.1	70.2
GDSTrack [20]	IJCAI 25	Self	45.9	39.3	35.4	70.9	48.5	73.9	59.8
SSTrack [54]	AAAI 25	Self	54.8	49.8	38.4	78.5	55.8	79.4	65.3
LGCTrack(Ours)	-	Self	59.5	55.3	45.7	84.7	62.0	91.9	77.6

In particular, compared with GDSTrack [20], LGCTrack yields absolute improvements of 13.6% PR, 16.0% NPR, and 10.3% SR. It also outperforms SSTrack [54] by 4.7%, 5.5%, and 7.3% on the three metrics, respectively. While fully supervised trackers still hold an advantage on this challenging large-scale benchmark, these consistent gains over prior self-supervised approaches clearly demonstrate the effectiveness of our learning framework.

RGBT234. RGBT234 [17] comprises 234 RGB-T video pairs and provides 12 challenge attributes for fine-grained evaluation. As reported in Tab. 1, LGCTrack achieves 84.7% PR and 62.0% SR, achieving the best performance among all self-supervised trackers. Compared with SSTrack [54], LGCTrack yields absolute gains of 6.2% in both PR and SR. Relative to GDSTrack [20], the improvements further increase to 13.8% in PR and 13.5% in SR. Moreover, LGCTrack remains competitive with several fully supervised trackers, further demonstrating the effectiveness of the proposed representation learning strategy.

GTOT. GTOT [16] is a widely adopted RGB-T benchmark consisting of 50 sequences collected in diverse scenarios, including offices, public roads, and waterside environments. On this dataset, LGCTrack achieves 91.9% PR and 77.6% SR, establishing the best performance among existing self-supervised methods

Table 2: Attribute-based comparison on RGBT234, where "LI", "LR", "NO", "TC", and "BC" denote low illumination, low resolution, no occlusion, thermal crossover, and background clutter, respectively.

Method	LI		LR		NO		TC		BC	
	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑
GDSTrack [20]	69.8	46.6	62.0	38.8	79.3	56.6	72.9	49.5	64.3	40.6
SSTrack [54]	67.5	46.4	75.3	49.6	86.7	64.3	84.8	60.7	65.1	42.7
LGCTrack	85.3	61.9	85.1	59.9	94.5	72.6	87.2	64.5	79.0	54.8

by a significant margin. In particular, relative to SSTrack [54], LGCTrack yields improvements of 12.5% in PR and 12.3% in SR. Furthermore, despite relying on no per-frame supervision, LGCTrack slightly surpasses the fully supervised CAFormer [45] on both metrics. These results indicate that the proposed method can substantially alleviate the reliance on dense manual annotations while retaining highly competitive tracking performance.

Attribute-based Analysis. Tab. 2 reports the attribute-based evaluation on RGBT234. LGCTrack attains the best PR and SR across all evaluated attributes, showing that its performance gains are consistent under diverse challenge factors. Notably, the improvements are most pronounced on LI, LR, and BC, suggesting that the advantage of LGCTrack is particularly evident in scenarios with degraded illumination, limited appearance details, and strong background interference. Meanwhile, the method also maintains clear gains on NO and TC, indicating that its superiority is not confined to a few difficult cases but remains stable across different tracking conditions.

4.3 Ablation Study

Ablation of the LGCTrack Framework. We evaluate the contributions of the proposed HLC and TGC modules on LasHeR [18] and RGBT234 [17]. As reported in Tab. 3, each module improves the baseline, and their combination consistently achieves the best performance on both benchmarks. These results indicate that LGCTrack benefits from effective cross-modal collaboration and stable self-supervised optimization. Specifically, HLC exploits complementary RGB and thermal cues through reliability-aware interaction to learn more discriminative representations, while TGC stabilizes the learning process by providing more reliable target-state guidance at the turning phase. Together, they form an effective solution for self-supervised RGB-T tracking.

Ablation of Cross-Modal Interaction Designs. We further analyze the design of HLC in Tab. 4. Vanilla fusion already improves the baseline, showing the benefit of cross-modal interaction. Replacing it with hierarchical collaboration

Table 3: Ablation study of the proposed components. We compare the baseline with variants that add only HLC, only TGC, and both components.

HLC TGC	LasHeR			RGBT234		
	PR↑	NPR↑	SR↑	PR↑	SR↑	
	54.8	49.8	38.4	78.5	55.8	
✓	58.3	54.1	43.9	81.4	59.7	
✓	55.7	51.0	42.6	81.3	57.4	
✓ ✓	59.5	55.3	45.7	84.7	62.0	

Table 5: Effect of different HLC integration depths. We evaluate the module across shallow (1-4), middle (5-8), and deep (9-12) blocks, as well as all layers (1-12).

Depth	LasHeR			RGBT234		
	PR↑	NPR↑	SR↑	PR↑	SR↑	
1-4	59.6	55.6	46.5	81.8	58.4	
5-8	60.4	56.3	46.4	83.1	60.4	
9-12	57.8	53.3	44.6	82.9	60.5	
1-12	59.5	55.3	45.7	84.7	62.0	

Table 4: Ablation of cross-modal design. *Baseline* denotes independent single-modality inference. *Vanilla* denotes direct addition of final-layer features.

Model Variant	LasHeR			RGBT234		
	PR↑	NPR↑	SR↑	PR↑	SR↑	
Baseline	54.8	49.8	38.4	78.5	55.8	
Vanilla	58.3	53.3	41.7	80.8	57.3	
HLC w/o gate	58.2	54.2	44.4	83.0	60.4	
LGCTrack	59.5	55.3	45.7	84.7	62.0	

Table 6: Effect of different update strategies in the TGC module. *Dir.* directly exchanges RGB and TIR global tokens without TGC. *Self* applies TGC and writes each refined token back to its original stream. *Cross* applies TGC and writes each refined token to the other stream.

Exp.	Dir.	Self	Cross	LasHeR			RGBT234		
				PR↑	NPR↑	SR↑	PR↑	SR↑	
I	✓			58.9	54.6	44.1	82.6	60.8	
II		✓		58.4	54.3	45.7	82.9	60.2	
LGCTrack			✓	59.5	55.3	45.7	84.7	62.0	

brings further gains, suggesting that progressively exploiting modality complementarity is more effective. Moreover, the full HLC with dynamic gating performs best on both benchmarks, highlighting the importance of reliability-aware modulation. By adaptively controlling information exchange, the dynamic gate enables more effective cross-modal collaboration, leading to more robust target representations in challenging scenarios.

Effect of the HLC Integration Depth. Tab. 5 analyzes the effect of inserting HLC at different backbone depths. We evaluate HLC in shallow (1–4), middle (5–8), and deep (9–12) blocks, as well as across all layers (1–12). Overall, restricting HLC to a specific stage limits its gains, while applying it across the entire encoder achieves the best overall performance across benchmarks. This indicates that continuous cross-modal collaboration progressively refines complementary cues across semantic levels. Therefore, we adopt the 1–12 setting as default in subsequent experiments.

Effect of Different Update Strategies in TGC. We compare different update strategies in TGC at the turning phase. As shown in Tab. 6, the best

Table 7: Ablation of TGC update scope and stage. *To-TIR* and *To-RGB* update only the TIR or RGB global token. *F2B* applies TGC only at the forward-to-backward transition.

Exp.	To-TIR	To-RGB	F2B	LasHeR			RGBT234		
				PR↑	NPR↑	SR↑	PR↑	SR↑	SR↑
III	✓		✓	57.8	53.0	43.1	80.9	58.8	
IV		✓	✓	56.3	51.4	41.5	78.0	56.7	
V	✓	✓		59.1	54.9	45.1	81.9	59.8	
LGCTrack	✓	✓	✓	59.5	55.3	45.7	84.7	62.0	

Table 8: Comparison of model complexity and runtime speed. We report the number of parameters, FLOPs, and tracking speed in FPS.

Method	Params (M)	FLOPs (G)	FPS
Un-Track [42]	98.7	47.3	23.3
IPL [28]	176.6	243.2	18.2
STTrack [10]	128.3	181.2	42.4
GDSTrack [20]	202.1	129.4	15.0
LGCTrack	167.1	178.4	52.0

performance is achieved by updating the refined token to the other modality stream. Directly exchanging RGB and TIR global tokens is less effective, suggesting that simple token swapping is insufficient for cross-modal interaction at the turning phase. Writing the refined token back to its original stream also degrades performance. Overall, TGC works best when the refined token guides the other modality stream during backward tracking.

Effect of the Update Scope and Insertion Strategy in TGC. We further analyze the update scope and insertion strategy of TGC in Tab. 7. The results show that updating a single modality global token is consistently less effective than jointly updating both modalities, regardless of whether the updated token belongs to TIR or RGB. This indicates that effective global coordination requires both streams to participate, allowing complementary information to propagate across branches. Moreover, introducing TGC only between forward and backward tracking performs better than inserting it into the propagation process. Overall, TGC is most effective when jointly updating both modality global tokens at the forward-to-backward transition.

Complexity Comparison. Tab. 8 compares model complexity and runtime speed. Notably, LGCTrack achieves the highest tracking speed among all compared methods, running at 52.0 FPS, which clearly demonstrates its real-time capability. Meanwhile, its parameter count and FLOPs remain competitive, showing that LGCTrack delivers superior runtime efficiency under a comparable computational budget. These results highlight the clear runtime advantage of LGCTrack for real-time RGB-T tracking.

4.4 Qualitative Analysis.

Fig. 4 presents qualitative comparisons with recent self-supervised RGB-T trackers on the LasHeR benchmark. The results reflect a fundamental challenge in RGB-T tracking, where the reliability of RGB and thermal cues varies over time. In the first two rows, sudden thermal degradation causes GDSTrack [20]

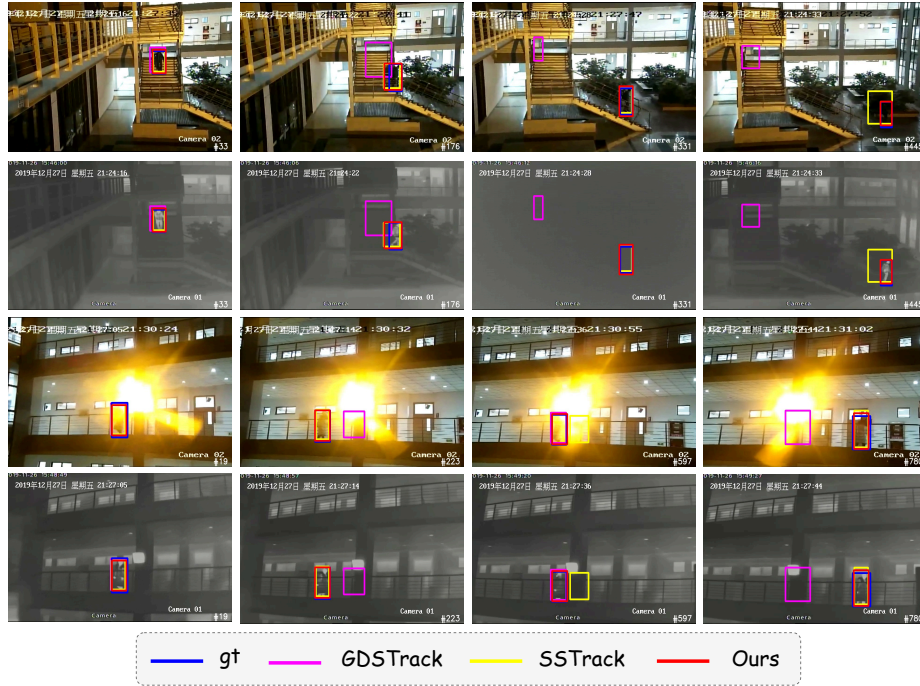


Fig. 4: Qualitative comparison of LGCTrack with self-supervised RGB-T trackers on the LasHeR. The results show that LGCTrack achieves more accurate and stable target localization under abrupt thermal degradation and severe RGB over-exposure.

to lose the target, while SSTrack [54] and LGCTrack remain effective. LGCTrack further produces more accurate and stable bounding boxes. In the last two rows, strong illumination severely over-exposes the RGB modality and changes the relative reliability of the two modalities. Under this condition, LGCTrack still maintains precise target localization throughout the sequence, whereas the competing methods show clear jitter and drift. These results demonstrate that LGCTrack adapts more effectively to dynamic modality degradation and exploits cross-modal complementarity more reliably for robust tracking.

5 Conclusion

This paper presents LGCTrack, a self-supervised RGB-T tracking framework for stable closed-loop learning under dynamic cross-modal reliability. LGCTrack introduces a local-to-global coordination strategy that enhances layer-wise feature interaction via Hierarchical Local Coordination and stabilizes target semantics at the forward-to-backward transition via Turning-phase Global Coordination. This design mitigates error accumulation and modality drift in closed-loop verification, enabling more reliable optimization. Extensive experiments show that

LGCTrack consistently outperforms existing self-supervised trackers and substantially narrows the gap to fully supervised methods. Overall, effective RGB-T tracking benefits from coordinated cross-modal interaction throughout the tracking cycle beyond cross-modal fusion alone.

Acknowledgements

This work was supported in part by Science and Technology Innovation Program of Xiongan New Area (Grant 2025XAGG0039), in part by the National Natural Science Foundation of China (Grants 62436002 and 62476198), in part by the Tianjin University Kunpeng&Ascend Center of Cultivation, in part by Tianjin Natural Science Funds for Distinguished Young Scholar (Grant 23JCJQJC00270), and in part by the Natural Science Foundation of Tianjin (Grant 25JCYBJC00950).

References

1. Cao, B., Guo, J., Zhu, P., Hu, Q.: Bi-directional adapter for multimodal tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 927–935 (2024)
2. Chan, S., Li, Z., Li, W., Lu, S., Shen, C., Zhang, X.: Smstracker: Tri-path score mask sigma fusion for multi-modal tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4766–4775 (2025)
3. Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W., Ouyang, W.: Backbone is all your need: A simplified architecture for visual object tracking. In: European conference on computer vision. pp. 375–392. Springer (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)
6. Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J.: Aiatracker: Attention in attention for transformer visual tracking. In: European conference on computer vision. pp. 146–164. Springer (2022)
7. Gao, Z., Li, D., Kuai, Y., Chen, R., Wen, G.: Visible-infrared image alignment for uavs: Benchmark and new baseline. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025)
8. 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. pp. 19079–19091 (2024)
9. Hou, X., Xing, J., Qian, Y., Guo, Y., Xin, S., Chen, J., Tang, K., Wang, M., Jiang, Z., Liu, L., et al.: Sdstracker: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26551–26561 (2024)

10. 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. vol. 39, pp. 3581–3589 (2025)
11. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence* **43**(5), 1562–1577 (2019)
12. Hui, T., Xun, Z., Peng, F., Huang, J., Wei, X., Wei, X., Dai, J., Han, J., Liu, S.: Bridging search region interaction with template for rgb-t tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13630–13639 (2023)
13. Kang, B., Chen, X., Lai, S., Liu, Y., Liu, Y., Wang, D.: Exploring enhanced contextual information for video-level object tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 4194–4202 (2025)
14. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
15. 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 Transactions on Pattern Analysis and Machine Intelligence* **47**(1), 634–649 (2024)
16. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing* **25**(12), 5743–5756 (2016)
17. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition* **96**, 106977 (2019)
18. 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 Transactions on Image Processing* **31**, 392–404 (2021)
19. Li, S., Yao, R., Zhou, Y., Zhu, H., Liu, B., Zhao, J., Shao, Z.: Unsupervised rgb-t object tracking with attentional multi-modal feature fusion. *Multimedia Tools and Applications* **82**(15), 23595–23613 (2023)
20. Li, S., Yao, R., Zhou, Y., Zhu, H., Sun, K., Liu, B., Shao, Z., Zhao, J.: Modality-guided dynamic graph fusion and temporal diffusion for self-supervised rgb-t tracking. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 1422–1430 (2025)
21. Li, S., Yao, R., Zhou, Y., Zhu, H., Zhao, J., Shao, Z., El Saddik, A.: Motion-aware self-supervised rgbt tracking with multi-modality hierarchical transformers. *ACM Transactions on Multimedia Computing, Communications and Applications* **20**(12), 1–23 (2024)
22. Li, X., Pei, W., Wang, Y., He, Z., Lu, H., Yang, M.H.: Self-supervised tracking via target-aware data synthesis. *IEEE transactions on neural networks and learning systems* **35**(7), 9186–9197 (2023)
23. Lin, L., Fan, H., Zhang, Z., Wang, Y., Xu, Y., Ling, H.: Tracking meets lora: Faster training, larger model, stronger performance. In: European Conference on Computer Vision. pp. 300–318. Springer (2024)
24. Liu, L., Li, C., Tang, J., Li, C.: Rgbt tracking via supervised mutual guiding. *Pattern Recognition* p. 112295 (2025)
25. Liu, L., Li, C., Xiao, Y., Ruan, R., Fan, M.: Rgbt tracking via challenge-based appearance disentanglement and interaction. *IEEE Transactions on Image Processing* **33**, 1753–1767 (2024)

26. 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. pp. 3129–3137 (2023)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
28. Lu, A., Li, C., Zhao, J., Tang, J., Luo, B.: Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks. *International Journal of Computer Vision* **133**(5), 2599–2619 (2025)
29. Lu, A., Wang, W., Li, C., Tang, J., Luo, B.: Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5793–5801 (2025)
30. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). pp. 300–317 (2018)
31. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
32. 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. pp. 22089–22098 (2025)
33. Sio, C.H., Ma, Y.J., Shuai, H.H., Chen, J.C., Cheng, W.H.: S2siamfc: Self-supervised fully convolutional siamese network for visual tracking. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1948–1957 (2020)
34. Tan, Y., Shao, J., Zamfir, E., Li, R., An, Z., Ma, C., Paudel, D., Van Gool, L., Timofte, R., Wu, Z.: What you have is what you track: Adaptive and robust multimodal tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3455–3465 (2025)
35. 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. pp. 5734–5744 (2025)
36. Tang, Z., Xu, T., Wu, X.J., Zhu, X., Cheng, C., Feng, Z., Kittler, J.: Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution. *IEEE Transactions on Image Processing* (2025)
37. 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. vol. 38, pp. 5189–5197 (2024)
38. 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. vol. 38, pp. 5436–5444 (2024)
39. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1308–1317 (2019)
40. Wang, N., Zhou, W., Song, Y., Ma, C., Liu, W., Li, H.: Unsupervised deep representation learning for real-time tracking. *International Journal of Computer Vision* **129**(2), 400–418 (2021)
41. Wu, Q., Wan, J., Chan, A.B.: Progressive unsupervised learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2993–3002 (2021)

42. Wu, Z., Zheng, J., Ren, X., Vasluianu, F.A., Ma, C., Paudel, D.P., Van Gool, L., Timofte, R.: Single-model and any-modality for video object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19156–19166 (2024)
43. Xia, J., Shi, D., Song, K., Song, L., Wang, X., Jin, S., Zhao, C., Cheng, Y., Jin, L., Zhu, Z., et al.: Unified single-stage transformer network for efficient rgb-t tracking. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 1471–1479 (2024)
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. vol. 36, pp. 2831–2838 (2022)
45. 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. vol. 39, pp. 8682–8690 (2025)
46. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporal transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19300–19309 (2024)
47. 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. pp. 3492–3500 (2022)
48. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European conference on computer vision. pp. 341–357. Springer (2022)
49. Yuan, D., Chang, X., Huang, P.Y., Liu, Q., He, Z.: Self-supervised deep correlation tracking. *IEEE Transactions on Image Processing* **30**, 976–985 (2020)
50. Zhang, P., Zhao, J., Bo, C., Wang, D., Lu, H., Yang, X.: Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing* **30**, 3335–3347 (2021)
51. 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. pp. 5404–5413 (2023)
52. Zhang, Z., Xu, L., Peng, D., Rahmani, H., Liu, J.: Diff-tracker: text-to-image diffusion models are unsupervised trackers. In: European Conference on Computer Vision. pp. 319–337. Springer (2024)
53. Zheng, J., Ma, C., Peng, H., Yang, X.: Learning to track objects from unlabeled videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13546–13555 (2021)
54. Zheng, Y., Zhong, B., Liang, Q., Li, N., Song, S.: Decoupled spatio-temporal consistency learning for self-supervised tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10635–10643 (2025)
55. Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X.: Odtrack: Online dense temporal token learning for visual tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7588–7596 (2024)
56. 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. pp. 9516–9526 (2023)