RELIABLE IDENTITY MEMORY FOR RGB-T MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-thermal multi-object tracking aims to localize multiple targets and preserve their identities across paired visible and thermal videos. Despite effectively fusing multimodal information, existing methods struggle to maintain consistent identity associations over time.This challenge is particularly acute for tiny objects, whose weak spatial evidence often leads to intermittent detections, disrupting temporal continuity and preventing reliable identity association. To overcome this challenge, we introduce Reliable Identity Memory Tracking (RIMTracking), a joint detection-and-tracking framework that preserves identity consistency by updating trajectory memory only with reliability-screened observations. Specifically, we employ an Observation-Filtered Identity Memory (OFIM) module to recursively consolidate cross-frame observations into a temporally evolving trajectory representation for stable identity association under changing modality reliability. We further develop Scale-Adaptive Cross-Modal Detail Refinement (SCDR), which sparsely samples fine-grained visible and thermal features and injects the aggregated evidence into object queries according to target scale, enhancing small-target localization. Extensive experiments and comparative analyses on VT-Tiny-MOT and VT-MOT demonstrate that our method achieves state-of-the-art performance while consistently improving identity preservation and tiny-object tracking.

1 INTRODUCTION

RGB-thermal (RGB-T) multi-object tracking organizes observations from paired visible and thermal videos into identity-consistent trajectories over time. The resulting trajectories support downstream tasks such as motion prediction and behavior analysis (Xu et al., 2024; Gao & Wang, 2023). Such capabilities are relevant to video surveillance and UAV-based monitoring, where targets must be followed under varying illumination (Zhu et al., 2025; Xu et al., 2024). By combining visible appearance details with complementary thermal cues, RGB-T tracking can provide observations when visible imagery alone is insufficient (Zhu et al., 2025). Reliable tracking of small targets under these conditions can therefore extend the use of trajectory-based video analysis to challenging low-light scenes.

Existing RGB-T multi-object trackers combine cross-modal information and temporal relations to improve object localization and trajectory association in challenging scenes (Zhu et al., 2025; Xu et al., 2024). Progressive fusion, heterogeneous graph modeling, and iterative cross-modal refinement allow these trackers to draw on complementary evidence from different modalities and neighboring frames (Zhu et al., 2025; Xu et al., 2024; Li et al., 2026). However, changes in illumination, occlusion, and target–background thermal contrast alter the information available from each modality, making a fixed fusion strategy difficult to apply across observation conditions (Xu et al., 2024; Ding et al., 2025). To accommodate such variation, RGB-T tracking research has also explored uncertainty-aware fusion that assesses modality reliability to reduce interference from unreliable inputs (Ding et al., 2025). For small objects, limited spatial information compounds this challenge: when a target occupies only a few pixels, the appearance cues and local details available for fusion may already be insufficient (Zhu et al., 2025; Xu et al., 2024). Even when a fused observation supports localization in the current frame, the tracker must still determine which historical trajectory the object belongs to. RGB-T tracking of small objects therefore requires both extracting useful observations from limited and changing multimodal evidence and using them to maintain identity continuity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

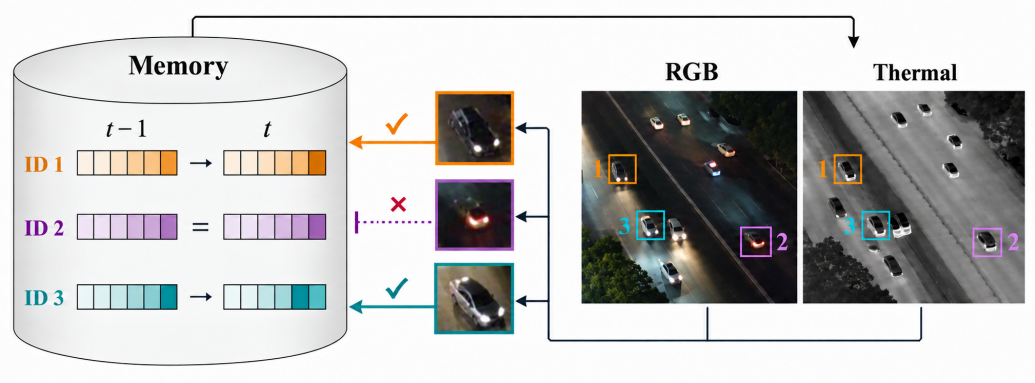


Figure 1: Identity association with selective memory updates. Historical memory provides context for associating current RGB-T targets. After identity assignment, reliable target features update memory (IDs 1 and 3), while unsupported updates are rejected and the existing state is preserved (ID 2). Colors denote identities, and feature bars illustrate memory states before and after the update. Scenes, boxes, and update decisions are illustrative.

To exploit history for identity association, learned association methods predict identities from current observations and historical trajectories, while memory-augmented trackers retain information over longer time horizons (Gao et al., 2025; Gao & Wang, 2023; Segu et al., 2025). The value of this history, however, depends not only on its temporal extent but also on the reliability of the observations it contains. Prior work has used uncertainty-based template selection in RGB-T single-object tracking and observation filtering in memory-based multi-object tracking to limit the influence of unreliable inputs on subsequent predictions (Ding et al., 2025; Segu et al., 2025). In multi-object tracking, reliability also concerns identity assignment: even a clearly localized object with a high detection score may be associated with the wrong trajectory. If such an observation updates that trajectory's long-term identity representation, a single association error may affect identity decisions in subsequent frames. For small objects, missing details and intermittent detections further restrict the evidence available for association, while filtering unreliable observations alone cannot recover information missing from the current frame. This calls for enhancing target details in weak observations while selecting evidence for identity memory updates according to both observation quality and association confidence. The central requirement is thus to adapt to changing RGB-T observations while selectively accumulating reliable trajectory information that remains useful for future identity association.

To address these challenges, we propose RIMTracking, an end-to-end RGB-T joint detection-and-tracking framework. As illustrated in Figure 1, its central idea is to distinguish using an observation for current identity prediction from retaining it in long-term memory. Specifically, we introduce the Observation-Filtered Identity Memory (OFIM) module, which maintains a persistent state for each trajectory and uses it to modulate historical ID embeddings for identity prediction. After identity assignment, OFIM evaluates observation quality and association confidence to determine whether the associated features should update this state, preserving the existing memory when an update is insufficiently supported.

Selective memory updates alone cannot compensate for the limited spatial evidence available for tiny objects. We therefore introduce Scale-Adaptive Cross-Modal Detail Refinement (SCDR) to supplement their current representations with fine-grained RGB-T information. Guided by object queries and their reference boxes, SCDR retrieves local details from auxiliary high-resolution visible and thermal features and adjusts their contribution according to target scale, placing greater emphasis on detail refinement for smaller objects.

Our main contributions are as follows:

- We propose RIMTracking, a joint detection-and-tracking framework for RGB-T multi-object tracking that distinguishes observation use in current identity prediction from information retention in long-term identity memory. Its core OFIM module provides per-trajectory memory context for identity association and selectively updates this memory after identity assignment based on observation quality and association confidence.

- We introduce SCDR, which uses object queries and their reference boxes to retrieve local details from auxiliary high-resolution RGB and thermal features. By modulating their contribution to each query according to the predicted object scale, SCDR supplements current observations of small objects with fine-grained spatial evidence.
- Experiments on VT-Tiny-MOT and VT-MOT show that RIMTracking achieves the best overall tracking performance among all compared methods on both benchmarks. These results demonstrate its effectiveness in both tiny-object tracking and the broader RGB-T tracking scenarios covered by VT-MOT.

2 Related Work

Association and persistent memory. Tracking-by-detection associates observations using motion, geometry, and appearance cues (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022; Aharon et al., 2022). Learned approaches also propagate track queries between frames (Zeng et al., 2022; Meinhardt et al., 2022), or formulate association as ID prediction from historical object features and identity embeddings (Gao et al., 2025). Memory-based trackers provide several ways to retain that context: MeMOT aggregates stored observations (Cai et al., 2022), MeMOTR maintains a slowly updated long-term representation (Gao & Wang, 2023), and Samba models tracklet sequences with synchronized states and masks uncertain observations (Segu et al., 2025). OFIM combines a persistent EMA state with historical-ID conditioning and screens updates after identity assignment, using identity confidence, temporal spacing, and box-scale consistency alongside detection confidence.

Multispectral detection and tracking. Multispectral detectors address feature misalignment through cross-modal alignment (Zhang et al., 2019), uncertainty-aware fusion (Sun et al., 2022), and transformer-based interaction (Xing et al., 2024; Yuan & Wei, 2024). MS-DETR sparsely samples the two modalities with loosely coupled attention and balances their contributions during optimization (Xing et al., 2024). For RGB-T tracking, PFTrack first integrates temporal information and then fuses modalities (Zhu et al., 2025), while HGT-Track connects detection and tracking nodes through heterogeneous spatial, temporal, and cross-modal relations (Xu et al., 2024). Our detector builds on multispectral feature extraction, while OFIM addresses a subsequent decision: whether an associated observation should modify the trajectory's persistent identity representation.

Fine-scale evidence for object queries. Feature pyramids preserve representations across scales (Lin et al., 2017), and deformable attention retrieves sparse features around object references (Zhu et al., 2021). QueryDet uses coarse small-object hypotheses to activate high-resolution computation (Yang et al., 2022). SCDR instead supplies fusion-decoder queries with an auxiliary stride-4 RGB-T feature path. Its size-dependent gate controls the contribution of retrieved detail to each query, while attention is still evaluated for all queries.

3 Method

3.1 Overview

In this section, we introduce RIMTracking, an end-to-end RGB-T multi-object tracker that combines cross-modal detail refinement with selective identity memory to enhance tiny-object representations and maintain identity consistency across frames. As illustrated in Figure 2, given paired visible and thermal video frames, the model first extracts and fuses cross-modal features to produce object representations and the corresponding detections in the current frame. These representations are then passed to an ID prediction module, which uses historical trajectory information to predict an existing identity or a newborn label for each object, thereby forming continuous trajectories. Building on this framework, we introduce Scale-Adaptive Cross-Modal Detail Refinement (SCDR) in Section 3.2 and Observation-Filtered Identity Memory (OFIM) in Section 3.3, and present the overall training objective in Section 3.4.

3.2 Scale-Adaptive Cross-Modal Detail Refinement

Tiny objects occupy only a few spatial locations in low-resolution feature maps, motivating the use of finer local information to refine their representations. To this end, we introduce SCDR, which

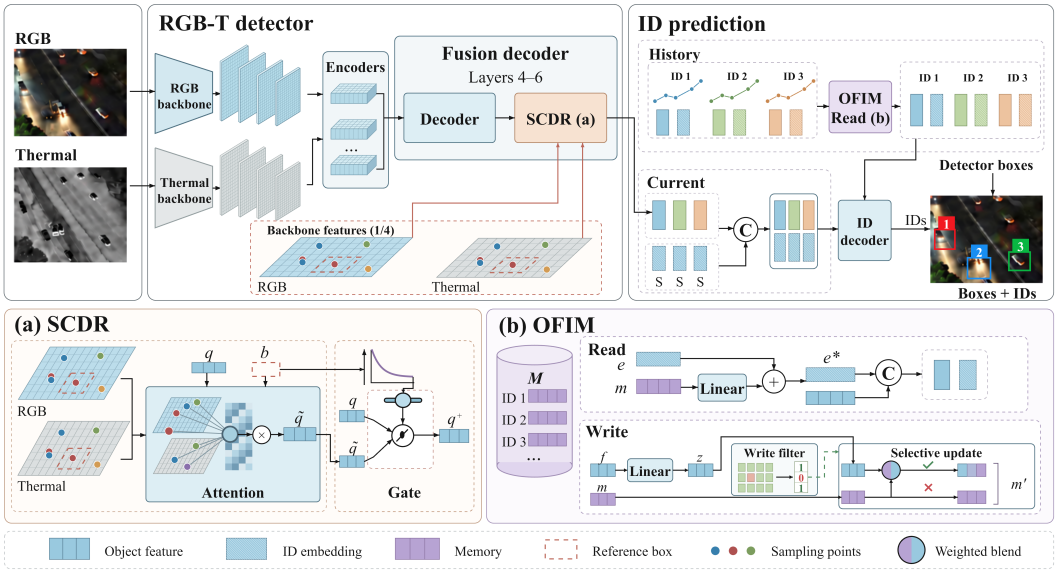


Figure 2: RIMTracking overview. (a) SCDR retrieves auxiliary RGB-T detail around query reference boxes and applies a scale-dependent residual. (b) OFIM conditions historical ID embeddings with persistent memory, then selectively writes assigned observations after identity prediction.

uses object queries to retrieve local details from high-resolution visible and thermal features and adjusts their contribution according to object scale.

Query-guided detail retrieval. Specifically, we extract stride-4 features from the visible and thermal backbones and project them to a shared channel dimension through modality-specific 1×1 convolutions and normalization, obtaining $\mathbf{P}_t^v$ and $\mathbf{P}_t^r$ at frame t. These feature maps serve as separate sources of detail for later decoder layers, while the original multi-scale encoder path remains unchanged. For the query $\mathbf{q}_{t,\ell}^i$ after the regular update in decoder layer ℓ, we use its positional embedding $\mathbf{p}_{t,\ell}^i$ and the reference box $\mathbf{b}_{t,\ell}^i$ entering that layer to retrieve additional information from both modality maps through deformable attention:

$$\mathbf{a}_{t,\ell}^i = \mathrm{MSDeformAttn}(\mathbf{q}_{t,\ell}^i + \mathbf{p}_{t,\ell}^i, \mathbf{b}_{t,\ell}^i, \{\mathbf{P}_t^v, \mathbf{P}_t^r\}). \tag{1}$$

The two modality maps are treated as separate attention levels, with sampling offsets and attention weights predicted from the query to aggregate local information guided by the reference box. The aggregated features are combined with the original query through residual connections, layer normalization, and a feed-forward network, yielding the refined representation $\widetilde{\mathbf{q}}_{t,\ell}^i$.

Scale-adaptive refinement. To adapt the contribution of local details to object scale, we compute $a_{t,\ell}^i = \sqrt{w_{t,\ell}^i h_{t,\ell}^i}$ from the normalized width $w_{t,\ell}^i$ and height $h_{t,\ell}^i$ of the reference box and map it to a continuous blending weight. Specifically, a sigmoid function determines the scale-dependent weight used to blend the refined representation with the original query:

$$g_{t,\ell}^i = \sigma\left(\frac{\tau_a - a_{t,\ell}^i}{T_a}\right), \qquad \mathbf{q}_{t,\ell}^{i,+} = \mathbf{q}_{t,\ell}^i + g_{t,\ell}^i(\widetilde{\mathbf{q}}_{t,\ell}^i - \mathbf{q}_{t,\ell}^i). \tag{2}$$

Here, σ denotes the sigmoid function, τ_a sets the transition scale, and T_a controls the smoothness of the transition. Smaller objects receive a higher blending weight for the retrieved details, whereas the updated queries remain closer to their original representations as object scale increases. The refined queries are then used to update the bounding boxes at the current layer and passed to subsequent decoding stages, ultimately producing object representations for identity prediction.

3.3 Observation-Filtered Identity Memory

SCDR supplements local details in current observations, while cross-frame identity association also requires the continuous accumulation of useful historical information. To this end, we introduce

OFIM, which maintains a persistent memory for each trajectory and separates memory reading for current identity prediction from selective updating after identity assignment.

Memory-conditioned identity prediction. At time t, we denote the memory of trajectory j before processing the current observation by $\mathbf{m}_{t^-}^j$, which retains information from previously accepted observations. Before identity prediction, we apply layer normalization and a learned projection to this memory and add the resulting residual to the historical ID embeddings:

$$\widehat{\mathbf{e}}_{t,k}^j = \mathbf{e}_{t,k}^j + \alpha W_p \,\mathrm{LN}(\mathbf{m}_{t^-}^j). \tag{3}$$

Here, k indexes historical positions, LN denotes layer normalization, W_p is a learned projection, and α is a learnable scalar controlling the memory contribution. All valid historical positions of the same trajectory share this memory residual while retaining their individual object features. We then concatenate the historical object features with the conditioned ID embeddings and pair current object features with the unknown-ID embedding, forming the inputs to the ID decoder for identity prediction. This reading process provides persistent identity information for current association, while the memory state remains unchanged until an update decision is made after identity assignment.

Selective memory update. After identity assignment, we define a binary memory-write indicator C_t^{ij} to determine whether observation i updates trajectory j's initialized memory. It equals 1 only when all observation-quality, identity-confidence, update-spacing, and scale-consistency conditions below are satisfied, and 0 otherwise:

$$\begin{aligned} C_t^{ij} = {} & [s_t^i \geq \tau_d] \wedge [p_t^i \geq \tau_p] \wedge [\Delta p_t^i \geq \tau_m] \\ & \wedge [t - u^j > \delta] \wedge [r_{\min} \leq A_t^i / \bar{A}^j \leq r_{\max}]. \end{aligned} \tag{4}$$

Here, s_t^i is the detection score, while p_t^i and Δp_t^i are the highest existing-ID score and the difference between the two highest existing-ID scores after applying association priors, excluding the newborn category. The frame of the last accepted write is denoted by u^j, and the minimum spacing δ limits the update frequency. Scale consistency is measured by the ratio of the current box area A_t^i to the median area $\bar{A}^j$ of recent valid historical boxes, rejecting observations whose area changes fall outside the prescribed range. For an accepted observation, we project its object feature as $\mathbf{z}_t^i = W_o \mathbf{o}_t^i$ and update the existing memory using an exponential moving average, following the long-term memory update in MeMOTR (Gao & Wang, 2023):

$$\mathbf{m}_t^j = \mathbf{m}_{t^-}^j + C_t^{ij} \lambda (\mathbf{z}_t^i - \mathbf{m}_{t^-}^j). \tag{5}$$

Here, λ controls the contribution of the new observation, and the persistent memory retains its previous value when the observation is rejected. An uninitialized memory is set directly to the first accepted projected observation, with newborn trajectories using an initialization criterion based on detection confidence. Thus, a current observation can participate in identity association without being written to persistent memory, while accepted updates provide identity information for subsequent frames.

3.4 Loss Function

We train RIMTracking with a combined objective comprising detection supervision, identity classification, and memory regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{id}} \mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{mem}} \mathcal{L}_{\mathrm{mem}}. \tag{6}$$

Here, $\mathcal{L}_{\mathrm{det}}$ comprises focal loss for object classification and L1 and GIoU losses for bounding box regression. $\mathcal{L}_{\mathrm{id}}$ is the cross-entropy loss for identity classification, supervising the identity predictions of current objects. $\mathcal{L}_{\mathrm{mem}}$ combines temporal contrast, identity separation, drift regularization, and norm regularization to promote temporal consistency and identity discrimination while limiting abrupt changes and excessive magnitudes in the learned representations. The coefficients λ_{id} and λ_{mem} balance the contributions of identity classification and memory regularization.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

Datasets. To evaluate RIMTracking on tiny objects and challenging RGB-T tracking scenarios, we use VT-Tiny-MOT and VT-MOT to assess tiny-object tracking and performance across platforms and scenes, respectively. VT-Tiny-MOT (Xu et al., 2024) contains 115 paired visible–thermal video sequences spanning seven object categories, with over 95.83% of target boxes smaller than 1,024 pixels in area. Its tiny objects, low illumination, and occlusion provide a suitable setting for evaluating detail refinement with SCDR and identity preservation with OFIM. VT-MOT (Zhu et al., 2025) comprises 582 paired sequences of pedestrians and vehicles captured by UAVs, surveillance cameras, and handheld devices under varied lighting and occlusion conditions. Together, these benchmarks enable us to evaluate RIMTracking under limited visual detail and challenges to identity continuity.

Metrics. Following prior MOT work (Gao et al., 2025; Segu et al., 2025), we use HOTA (Luiten et al., 2021) and IDF1 (Ristani et al., 2016) as our primary metrics, measuring overall detection–association quality and identity consistency, respectively. For completeness, we report MOTA and MOTP (Bernardin & Stiefelhagen, 2008) to measure tracking errors and localization precision, and additionally report DetA on VT-MOT. On VT-Tiny-MOT, we also report MT, ML, FP, FN, and IDSW to characterize trajectory coverage, false positives, missed detections, and identity switches. Since OFIM targets identity consistency, its ablations focus on HOTA, IDF1, AssA, and IDSW to assess how memory updates affect overall tracking and identity association.

4.2 IMPLEMENTATION DETAILS

RIMTracking builds upon the identity prediction framework of MOTIP (Gao et al., 2025) and adopts MS-DETR (Xing et al., 2024) for RGB-T feature extraction and fusion. We use dual ResNet-50 backbones (He et al., 2016) and initialize the corresponding network parameters with Deformable DETR (Zhu et al., 2021) weights pretrained on COCO (Lin et al., 2014). We use 300 detection queries and a six-layer ID decoder, with object features, ID embeddings, and OFIM memory representations all set to 256 dimensions. We apply SCDR to layers 4–6 of the fusion decoder, using four sampling points per attention head and modality on stride-4 feature maps, with the scale-gating parameters $\tau_a = 0.05$ and $T_a = 0.01$. For OFIM, we set the EMA update coefficient to 0.01 and the update cooldown interval to five frames.

Training proceeds in three stages: detection pretraining, joint detection and tracking training, and fine-tuning of the ID decoder and OFIM. We train the tracking model using clips of 16 consecutive frames. Input images are resized to a target shorter-side length of 512 pixels, with the longer side capped at 640 pixels while preserving the aspect ratio. We apply horizontal flipping, random cropping, and resizing, together with trajectory occlusion and identity switching, with both trajectory augmentation probabilities set to 0.5. We use the AdamW optimizer with a base learning rate of 10^{-4} and a learning rate of 3×10^{-5} for the final fine-tuning stage.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

VT-Tiny-MOT. The limited appearance cues of tiny objects, together with degraded observations under low illumination and occlusion, pose substantial challenges to reliable detection and identity association in RGB-T videos. Although methods combining ProbEn or UA-CMDet with SORT exploit complementary visible and thermal information, they still fall behind RIMTracking in overall tracking performance and identity matching accuracy. As shown in Table 1, RIMTracking outperforms HGT-Track, the strongest RGB-T baseline in this comparison, by 6.808 HOTA, 4.362 IDF1, and 4.451 MOTA points under visible annotations. Meanwhile, false negatives decrease from 109,500 to 98,365, and MOTP increases from 55.6 to 66.541, indicating fewer missed targets and more accurate localization of matched detections. The number of mostly tracked trajectories increases from 226 to 283, while the number of mostly lost trajectories decreases from 444 to 351, indicating improved trajectory coverage for tiny objects. These results demonstrate the potential of RIMTracking for RGB-T tiny-object tracking by improving both target detection and identity matching in challenging scenes.

Table 1: VT-Tiny-MOT with visible annotations. Best values are bold.

Method	Mod.	Backbone	HOTA↑	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDF1↑
DeepSORT (Wojke et al., 2017)	V	ResNet-50	23.4	12.4	68.8	15	623	7043	148453	20.7
Tracktor (Bergmann et al., 2019)	V	ResNet-50	23.3	3.0	64.2	152	605	28494	144647	22.5
ByteTrack (Zhang et al., 2022)	V	ResNet-50	26.3	13.3	68.7	152	625	6670	148259	26.3
OCSORT (Cao et al., 2023)	V	ResNet-50	25.8	10.6	68.0	156	612	11940	146690	25.7
MPNTrack* (Braso & Leal-Taixe, 2020)	V	ResNet-50	23.0	3.5	62.4	168	594	29528	143452	24.6
DSFNet (Xiao et al., 2022)	V	DLA-34	16.3	10.4	64.3	32	747	**911**	159954	16.6
CenterTrack (Zhou et al., 2020)	V	DLA-34	11.8	10.3	**71.6**	16	680	1859	154977	8.8
FairMOT (Zhang et al., 2021)	V	DLA-34	22.7	12.1	67.6	76	688	2721	153421	22.4
TraDes (Wu et al., 2021)	V	DLA-34	20.0	8.8	66.0	82	691	9938	153532	20.4
GSDT* (Wang et al., 2021)	V	DLA-34	22.7	8.9	71.3	76	688	2722	153415	22.4
Transcenter (Xu et al., 2023)	V	PVTv2	5.0	-4.0	46.6	11	856	10522	176028	2.1
HGT-Track-V (Xu et al., 2024)	V	PVTv2	23.1	27.2	55.0	156	430	16860	110652	33.6
ProbEn-E+SORT (Chen et al., 2022)	V+T	ResNet-50	20.0	9.5	66.7	138	542	19024	134952	17.4
ProbEn-M+SORT (Chen et al., 2022)	V+T	ResNet-50	26.3	24.7	67.4	211	450	12108	111365	26.1
UA-CMDet+SORT (Sun et al., 2022)	V+T	ResNet-50	24.1	8.6	65.1	190	464	36612	121063	23.2
HGT-Track (Xu et al., 2024)	V+T	PVTv2	29.1	32.0	55.6	226	444	11594	109500	43.2
RIMTracking (ours)	V+T	ResNet-50	**35.908**	**36.451**	66.541	**283**	**351**	13075	**98365**	**47.562**

Table 2: VT-MOT Protocol I. Best values are bold.

Method	Modality	HOTA↑	DetA↑	MOTP↑	IDF1↑	MOTA↑
FairMOT (Zhang et al., 2021)	V+T	37.35	34.628	72.525	45.795	37.266
CenterTrack (Zhou et al., 2020)	V+T	39.045	38.104	72.874	44.42	30.585
TraDes (Wu et al., 2021)	V+T	38.319	36.37	72.295	47.008	34.632
TransTrack (Sun et al., 2020)	V+T	38.00	35.711	73.823	43.567	36.156
ByteTrack (Zhang et al., 2022)	V+T	38.393	32.122	73.483	45.757	33.151
OC-SORT (Cao et al., 2023)	V+T	31.479	25.244	73.15	38.086	28.948
MixSort-OC (Cui et al., 2023)	V+T	39.09	33.109	73.632	45.799	31.33
MixSort-Byte (Cui et al., 2023)	V+T	39.575	34.806	73.049	46.367	31.593
PID-MOT (Lv et al., 2024)	V+T	35.621	33.245	71.794	42.43	33.333
Hybrid-SORT (Yang et al., 2024)	V+T	39.485	34.619	72.840	46.310	31.074
PFTrack (Zhu et al., 2025)	V+T	41.068	41.631	**73.949**	47.254	43.088
RIMTracking (ours)	V+T	**45.456**	**44.474**	73.414	**56.554**	**48.578**

VT-MOT. Videos captured by drones, surveillance cameras, and handheld devices exhibit diverse viewpoints, accompanied by variations in object scale and illumination. PFTrack exploits complementary visible and thermal information through progressive fusion and achieves the highest HOTA and IDF1 among the existing methods listed in Table 2. RIMTracking further improves tracking performance, achieving the best HOTA, DetA, IDF1, and MOTA among all compared methods. In particular, it achieves 45.456 HOTA and 56.554 IDF1, exceeding PFTrack by 4.388 and 9.300 points, respectively, and demonstrating better overall tracking performance and identity matching accuracy. This improvement in identity matching is consistent with the design of OFIM, which selectively updates identity memory to limit the influence of unreliable observations on historical identity representations and provide persistent identity cues for current predictions. Meanwhile, DetA and MOTA reach 44.474 and 48.578, exceeding PFTrack by 2.843 and 5.490 points, respectively, indicating that improved identity matching is accompanied by gains in detection accuracy and overall tracking accuracy.

4.4 Ablation Studies

We conduct ablation experiments on VT-Tiny-MOT using visible annotations to assess the contribution of each component to RGB-T tiny-object tracking, the primary focus of this work. As shown in Table 3, we progressively add SCDR and OFIM to the tracking framework with the ID decoder; we also compare MS-DETR with and without SCDR, using BoT-SORT to track the resulting detections, to further examine the effect of detail refinement on tracking performance. We then keep the model weights and all other inference settings fixed and vary only the conditions for updating

Table 3: Component ablation on VT-Tiny-MOT using visible annotations. Compared to prior work (in gray), we introduce SCDR to refine cross-modal details and OFIM to selectively update identity memory. We progressively evaluate both modules in identity-prediction-based tracking and further assess SCDR with BoT-SORT. Best results within each ablation group are bold.

Method / association		SCDR	OFIM	HOTA↑	MOTP↑	IDF1↑	MOTA↑
ID decoder	(a)	–	–	34.980	65.535	45.441	35.252
	(b)	✓	–	35.504	66.137	45.848	35.393
	(c)	✓	✓	**35.908**	**66.541**	**47.562**	**36.451**
BoT-SORT (Aharon et al., 2022)	(d)	–	–	32.628	61.949	44.618	32.120
	(e)	✓	–	**34.134**	**63.448**	**47.174**	**33.230**
UA-CMDet+SORT (Sun et al., 2022)	(f)	–	–	24.1	65.1	23.2	8.6
HGT-Track (Xu et al., 2024)	(g)	–	–	29.1	55.6	43.2	32.0

Table 4: Write-policy ablation on VT-Tiny-MOT sequences. All settings except existing-memory write criteria are fixed. Full OFIM combines detection, identity, cooldown, and scale checks. Best values are bold.

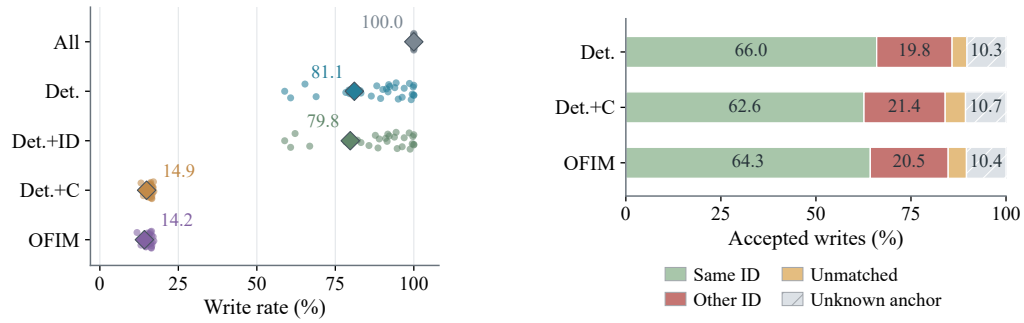
Write policy	HOTA↑	AssA↑	IDF1↑	IDSW↓
Unfiltered EMA	35.753	47.795	47.130	3399
Detection-filtered EMA	35.822	47.936	47.380	3370
Detection + identity filtering	35.725	47.668	47.198	3454
Detection + cooldown	35.903	48.128	47.507	3238
OFIM (full)	**35.908**	**48.142**	**47.562**	**3172**

existing identity memory to evaluate the effect of different write policies on identity association, as shown in Table 4.

Component effectiveness. As shown in Table 3, adding SCDR to the ID decoder baseline (row b) improves HOTA, MOTP, and IDF1 by 0.524, 0.602, and 0.407 points over row a, respectively. These results show that refining the detailed representations of tiny objects improves localization and provides more effective observation features for identity association. With OFIM, the full configuration (row c) further increases IDF1 to 47.562, a gain of 1.714 points over the configuration with SCDR alone, while HOTA rises from 35.504 to 35.908. This improvement is consistent with the role of OFIM in preserving historical identity information to support current identity predictions, reflecting further gains in identity association with the full configuration. When MS-DETR is combined with BoT-SORT, adding SCDR to MS-DETR (row e) likewise improves HOTA and IDF1 by 1.506 and 2.556 points over row d, respectively, demonstrating that detail refinement benefits different tracking methods. Our full model (row c) outperforms HGT-Track (row g) and UA-CMDet+SORT (row f) by 4.362 and 24.362 IDF1 points, respectively.

Memory write policy. Table 4 compares different memory-write policies, with unfiltered EMA accepting all candidate updates to existing memory as the baseline for selective writing. Filtering updates by detection confidence alone increases AssA and IDF1 from 47.795 and 47.130 to 47.936 and 47.380, respectively, indicating that observation filtering improves identity association. However, adding identity filtering to detection filtering does not yield further gains and instead increases IDSW from 3,370 to 3,454, showing that additional filtering criteria alone do not guarantee better tracking performance. In contrast, adding a cooldown interval to detection filtering increases AssA and IDF1 to 48.128 and 47.507 and reduces IDSW to 3,238, highlighting the benefit of spacing memory updates for identity association. Full OFIM, which combines detection, identity, cooldown, and scale checks, achieves the best results on all four metrics and further reduces identity switches by 66 compared with detection filtering with cooldown. Compared with unfiltered EMA, full OFIM reduces identity switches by 227 (approximately 6.7%), demonstrating that selectively accepting observations and controlling update timing better preserves identity continuity than unconditional updating.

Memory-write behavior. Figure 3 shows the update frequencies and the composition of accepted observations across sequences for different write policies, providing further insight into the memory-



(a) Frequency across 29 sequences

(b) Composition of accepted writes

Figure 3: Memory-write behavior under different policies on VT-Tiny-MOT. (a) Write rates across sequences, with dots and diamonds denoting per-sequence and pooled rates. (b) Composition of accepted observations relative to the ground-truth identity matched at initialization. All: unfiltered; Det.: detection filtering; ID: identity filtering; C: cooldown.

update behavior of OFIM. Excluding memory initialization, OFIM accepts only 14.2% of candidate updates, substantially fewer than detection-only filtering (81.1%) and close to detection filtering with cooldown (14.9%). Among accepted writes with a known identity anchor at initialization, the proportion of observations assigned to a different ground-truth identity or left unmatched is 28.2% for OFIM, lower than the 29.9% of detection filtering with cooldown but higher than the 26.5% of detection-only filtering. Together with Table 4, these results show that OFIM achieves better association performance with fewer memory updates, while its gains cannot be explained solely by the proportion of accepted observations consistent with the anchored identity. Figure 4 further illustrates the tracking behavior in nighttime and daytime examples. Detection-filtered EMA exhibits identity switches, whereas OFIM maintains consistent identities in the displayed clips.



(a) Nighttime: DJI_0073_3

(b) Daytime: DJI_0141_2

Figure 4: Tracking examples with detection-filtered EMA (Det.) and OFIM using identical crops. Numbers indicate frame indices and predicted IDs. Event strips show post-assignment actions for observations matched to the target ground truth; dashed lines mark displayed frames.

5 Conclusion

We introduce RIMTracking, which combines cross-modal detail refinement with selective identity memory to improve object representations and identity association for RGB-T tiny-object tracking. The proposed method exploits complementary visual information while retaining historical identity information when observations are unreliable, supporting more stable identity association. RIMTracking outperforms existing state-of-the-art methods on RGB-T multi-object tracking benchmarks, with further improvements in association accuracy and identity stability.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 941–951, 2019.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. doi:10.1155/2008/246309.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pp. 3464–3468, 2016.

Guillem Braso and Laura Leal-Taixe. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6247–6257, 2020.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. MeMOT: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8090–8100, 2022.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pp. 139–158, 2022. doi:10.1007/978-3-031-20077-9_9.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9921–9931, 2023.

Zhaodong Ding, Chenglong Li, Shengqing Miao, and Jin Tang. Template-based uncertainty multimodal fusion network for RGBT tracking. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 909–917. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi:10.24963/ijcai.2025/102.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9901–9910, 2023. doi:10.1109/ICCV51070.2023.00908.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as ID prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27883–27893, 2025. doi:10.1109/CVPR52734.2025.02596.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Weiran Li, Yeqiang Liu, Yijie Wei, Mina Han, Qiannan Guo, and Zhenbo Li. DM^{3}T: Harmonizing modalities via diffusion for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8398–8407, 2026. URL https://vranlee.github.io/DM-3-T/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548–578, 2021. doi:10.1007/s11263-020-01375-2.

W. Lv, N. Zhang, J. Zhang, and D. Zeng. One-shot multiple object tracking with robust ID preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4473–4488, 2024.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8844–8854, 2022.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops*, pp. 17–35, 2016.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations*, 2025.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. TransTrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi:10.1109/TCSVT.2022.3168279.

Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *IEEE International Conference on Robotics and Automation*, pp. 13708–13715, 2021.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, pp. 3645–3649, 2017.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12352–12361, 2021.

C. Xiao, Q. Yin, X. Ying, R. Li, S. Wu, M. Li, L. Liu, W. An, and Z. Chen. DSFNet: Dynamic and static fusion network for moving object detection in satellite videos. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi:10.1109/LGRS.2021.3124222.

Yinghui Xing, Shuo Yang, Song Wang, Shizhou Zhang, Guoqiang Liang, Xiuwei Zhang, and Yanning Zhang. MS-DETR: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20628–20642, 2024. doi:10.1109/TITS.2024.3450584.

Qingyu Xu, Longguang Wang, Weidong Sheng, Yingqian Wang, Chao Xiao, Chao Ma, and Wei An. Heterogeneous graph transformer for multiple tiny object tracking in RGB-T videos. *IEEE Transactions on Multimedia*, 26:9383–9397, 2024. doi:10.1109/TMM.2024.3401548.

Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7820–7835, 2023.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2022.

Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-SORT: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6504–6512, 2024. doi:10.1609/aaai.v38i7.28471.

Maoxun Yuan and Xingxing Wei. C2Former: Calibrated and complementary transformer for RGB-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi:10.1109/TGRS.2024.3376819.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pp. 659–675, 2022. doi:10.1007/978-3-031-19812-0_38.

Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5126–5136, 2019. doi:10.1109/ICCV.2019.00523.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. doi:10.1007/s11263-021-01513-4.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pp. 1–21, 2022. doi:10.1007/978-3-031-20047-2_1.

Xingyi Zhou, Vladlen Koltun, and Philipp Krahenbuhl. Tracking objects as points. In *European Conference on Computer Vision*, pp. 474–490, 2020.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, Chengjie Gu, and Zhixiang Huang. Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognition*, 161:111330, 2025. doi:10.1016/j.patcog.2024.111330.