Collaborating Vision, Depth, and Thermal Signals for Multi-Modal Tracking: Dataset and Algorithm

Xue-Feng Zhu[1], **Tianyang Xu**[1,*], **Yifan Pan**[1], **Jinjie Gu**[1],
Xi Li[2], **Jiwen Lu**[3], **Xiao-Jun Wu**[1], **Josef Kittler**[4]
[1] Jiangnan University; [2] Zhejiang Univeristy; [3] Tsinghua University; [4] University of Surrey

Abstract

Existing multi-modal object tracking approaches primarily focus on dual-modal paradigms, such as RGB-Depth or RGB-Thermal, yet remain challenged in complex scenarios due to limited input modalities. To address this gap, this work introduces a novel multi-modal tracking task that leverages three complementary modalities, including visible RGB, Depth (D), and Thermal Infrared (TIR), aiming to enhance robustness in complex scenarios. To support this task, we construct a new multi-modal tracking dataset, coined RGBDT500, which consists of 500 videos with synchronised frames across the three modalities. Each frame provides spatially aligned RGB, depth, and thermal infrared images with precise object bounding box annotations. Furthermore, we propose a novel multi-modal tracker, dubbed RDTTrack. RDTTrack integrates tri-modal information for robust tracking by leveraging a pretrained RGB-only tracking model and prompt learning techniques. In specific, RDTTrack fuses thermal infrared and depth modalities under a proposed orthogonal projection constraint, then integrates them with RGB signals as prompts for the pre-trained foundation tracking model, effectively harmonising tri-modal complementary cues. The experimental results demonstrate the effectiveness and advantages of the proposed method, showing significant improvements over existing dual-modal approaches in terms of tracking accuracy and robustness in complex scenarios. The dataset and source code are publicly available at `https://xuefeng-zhu5.github.io/RGBDT500`.

1 Introduction

Visual object tracking aims to automatically localise an object of interest within a video, based on the initially specified object position and scale [16, 24]. It is a fundamental research topic in the field of artificial intelligence and computer vision. Through decades of research, visual object tracking has witnessed substantial progress, driven by continuous advancements in benchmark datasets [14, 9, 56] and tracking algorithms [45, 46, 54]. However, conventional tracking methods mainly rely on visible RGB images, often suffer from reduced effectiveness and robustness under conditions of adverse visibility [37].

To cope with these challenges, recent research has explored the integration of an additional sensing modality to enhance tracking performance in complex and visually degraded scenarios. For instance, multi-modal tracking methods, such as RGB-D and RGB-T tracking, leverage complementary information from depth and thermal data, respectively, to improve robustness and accuracy in adverse conditions. In particular, RGB-D tracking [55, 57] combines an RGB image with a depth map to provide additional spatial and structural cues, enabling more accurate localisation in challenging scenarios such as occlusion, background clutter, and scale variation. Similarly, RGB-T tracking [38,

*Corresponding author: `tianyang.xu@jiangnan.edu.cn`

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

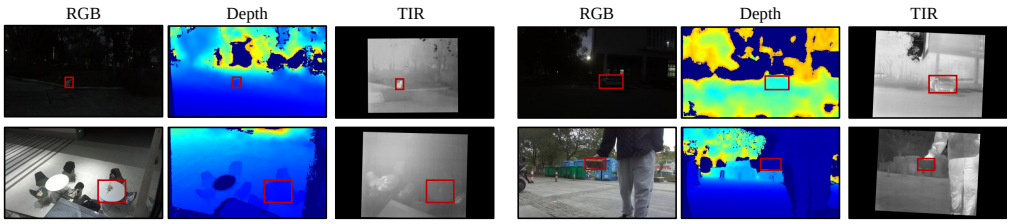


Figure 1: Representative tri-modal samples of RGBDT500. In each sample, at least one modality is affected by specific challenges, highlighting the need for more general multi-modal fusion.

44] integrates RGB and thermal infrared images, compensating for the limitations of visible light in low-illumination scenarios.

Despite these advantages, dual-modal RGB-D and RGB-T trackers still struggle to perform reliably in real-world environments characterised by simultaneous challenges such as poor lighting, occlusions, adverse weather, etc. As shown in Fig. 1, in some scenarios, both dual-modal RGB-T and RGB-D tracking frameworks may experience modality-specific information degradation due to factors such as thermal crossover, sensor noise, or depth ambiguity in complex scenes. To address these limitations, we construct RGBDT500, a comprehensive tri-modal tracking dataset comprising synchronised RGB, depth, and thermal infrared video sequences.

Specifically, RGBDT500 includes a total of 500 video sequences, with 400 used for training and 100 reserved for testing. Each frame in the dataset provides spatially aligned RGB, depth, and thermal infrared images. The training set of RGBDT500 consists of about 160K RGB-Depth-Thermal (RGB-D-T) image triplets, while the test set contains 43.7K RGB-D-T image triplets, each annotated with an object bounding box. Compared to existing dual-modal RGB-D and RGB-T tracking datasets, the three modalities of RGBDT500 offer richer information for object localisation. Based on the developed dataset, we naturally extend to a new multi-modal tracking task, tri-modal tracking, which requires leveraging three complementary modalities, including RGB, depth and thermal infrared, for robust tracking. By effectively integrating RGB, depth, and thermal infrared cues, tri-modal tracking enables adaptive enhanced performance in complex scenarios.

Although tri-modal tracking data provides more advantageous information, it also introduces additional challenges for current multi-modal tracking paradigms. Existing multi-modal tracking approaches primarily focus on fusing RGB with a single additional modality, making them inadequate for directly handling tri-modal data. This limitation is further compounded by the substantial differences among the three modalities. For tracking, RGB provides rich texture and colour details, thermal infrared highlights salient heat-emitting objects, while depth captures spatial and geometric structure. To address these challenges and validate the effectiveness of the RGBDT500 dataset, we propose a straightforward and effective baseline tracker, named RDTTrack.

In detail, to leverage the representation power of the pre-trained RGB tracking model and effectively integrate tri-modal information for object localisation, RDTTrack incorporates a prompt learning mechanism. Specifically, to fuse depth and thermal infrared cues, it utilises a feature projection to enforce orthogonality between depth and thermal infrared features. This orthogonal projection aims to reduce feature redundancy and enhance the reliability of the fused representation. Subsequently, fused depth and thermal features are integrated with RGB features as prompts for the pre-trained OSTrack [49] model, effectively leveraging tri-modal complementary cues for robust tracking. For RDTTrack training, the pre-trained OSTrack model is kept frozen, and only the prompt learning module is fine-tuned using the 400 training sequences from the RGBDT500 dataset. Extensive experiments, including ablation studies and comparative evaluations, demonstrate the effectiveness and the competitive performance of the proposed baseline tracker.

In summary, our contributions are as follows:

- We introduce a novel multi-modal object tracking task that incorporates three modalities to ensure robust tracking performance, thereby extending existing dual-modal tracking and promoting further progress in the field.

- We present RGBDT500, a dataset with temporally and spatially aligned RGB, depth, and thermal modalities, establishing a data foundation for advancing tri-modal tracking research.
- We propose a novel multi-modal baseline tracker, RDTTrack, which integrates prompt learning with orthogonality constraints to enable effective fusion of tri-modal information.
- Extensive experiments on the proposed RGBDT500 benchmark demonstrate the generalisation and effectiveness of our baseline RDTTrack.

2 Related Work

Recent visual tracking research has evolved from uni-modal to multi-modal approaches, enabling more robust performance in complex environments. This section provides an overview of research progress in multi-modal tracking, specifically focusing on relevant datasets, RGB-D and RGB-T tracking methodologies.

2.1 Multi-Modal Tracking Datasets

Developing multi-modal tracking datasets is essential for both advancing and rigorously evaluating tracking algorithms. To date, most existing datasets have concentrated on RGB-D and RGB-T modalities. Among RGB-D tracking datasets, DepthTrack [48] stands out as a significant benchmark, comprising 200 video sequences. More recently, the RGBD1K dataset [57] has been introduced, featuring 1,050 sequences, with object bounding box annotations provided for around 720K RGB-D frames. These large-scale datasets have been instrumental in advancing the development of RGB-D tracking approaches. Additionally, other important benchmarks, such as CDTB [30], PTB[36], and STC [43], are also valuable for evaluating RGB-D tracking algorithms.

In RGB-T tracking, several influential datasets have been introduced, including GTOT [19], RGBT234 [20], and LasHeR [21]. GTOT serves as the first benchmark dataset for RGB-T tracking, comprising 50 RGB-T video sequences. Building upon the GTOT benchmark, RGBT234 extends the scale to 234 RGB-T sequences, providing broader coverage of diverse tracking scenarios. To date, LasHeR represents the most extensive RGB-T tracking dataset, containing 1,244 RGB-T sequences and over 730K paired frames. The thermal modality, which is particularly effective under low-light and adverse illumination conditions, enhances tracking robustness in real-world environments. The release of these datasets has significantly contributed to the advancement of RGB-T tracking.

Despite these advancements, most existing multi-modal datasets are restricted to RGB-D or RGB-T modality combinations, lacking support for more general multi-modal tracking. Additionally, ensuring accurate temporal and spatial alignment between modalities and adequately covering the broad range of object categories across diverse tracking conditions remains a significant challenge. To address these gaps, we introduce RGBDT500, the first dataset featuring spatiotemporally aligned RGB, depth, and thermal modalities for visual tracking.

2.2 Multi-Modal Tracking Methodologies

RGB-D tracking leverages complementary information from both RGB and depth modalities, enabling more comprehensive scene understanding for tracking. A variety of trackers have been proposed to realise effective fusion and interaction between RGB and depth modalities. For example, DAL [32] integrates depth information into RGB features based on the correlation filter-based tracking framework. Recent trackers, such as DeT [48], adapt RGB-only tracking architectures [6, 2] by fusing feature maps through pixel-wise operations. SPT [57] utilises separated transformer encoders for RGB and depth, followed by a dedicated fusion module. ARKitTrack [52] further advances fusion by encoding depth into bird's eye view representations and performing cross-view fusion. Collectively, these approaches represent ongoing progress in RGB-D tracking, aiming to enhance multi-modal feature extraction, fusion, and adaptation for improved tracking robustness.

RGB-T tracking leverages the complementary strengths of RGB and thermal modalities to enhance tracking performance under challenging conditions such as low-light, nighttime environments. Early efforts primarily focus on convolutional neural network-based fusion strategies [39, 11]. For example, mfDiMP [50] introduces an end-to-end framework for RGB-T fusion, while MANet [18] employes a three-way adapter network to extract modality-specific and shared features between two modalities.

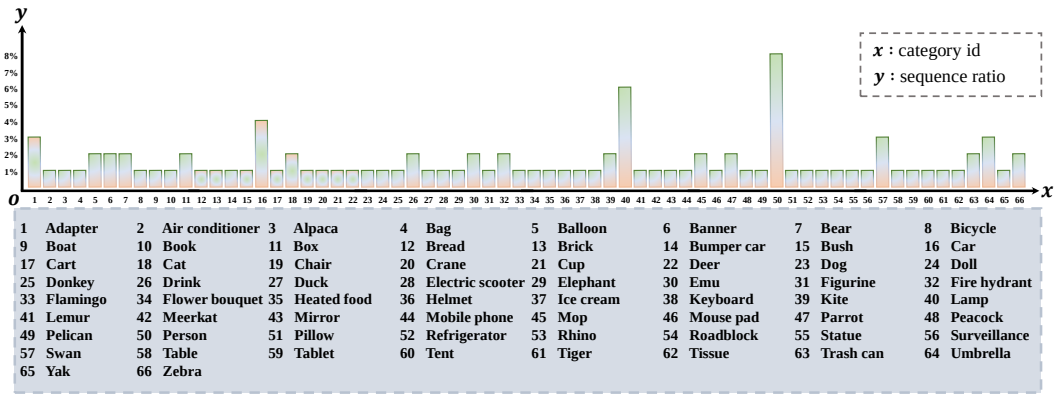


Figure 2: The object category distribution of the RGBDT500 test set.

Table 1: A comparison of RGBDT500 with related RGB-only, RGB-T, and RGB-D tracking datasets.

Dataset	Modalities			Sequences	Frames	Classes	Training Set	Publication
	RGB	Depth	TIR					
UAV123 [1]	✔	✘	✘	123	113K	-	✘	ECCV 2016
GOT-10k [14]	✔	✘	✘	10K	1.5M	563	✔	IEEE TPAMI 2019
LaSOT [9]	✔	✘	✘	1400	3.5M	70	✔	CVPR 2019
GTOT [19]	✔	✘	✔	50	15.8K	9	✘	IEEE TIP 2016
RGBT210 [22]	✔	✘	✔	210	104.7K	22	✘	ACM MM 2017
RGBT234 [20]	✔	✘	✔	234	116.7K	22	✘	PR 2019
LasHeR [21]	✔	✘	✔	1224	734.8K	32	✔	IEEE TIP 2021
VTUAV [51]	✔	✘	✔	500	1.7M	13	✔	CVPR 2022
PTB [36]	✔	✔	✘	100	21K	26	✘	CVPR 2013
CDTB [30]	✔	✔	✘	80	101.9K	21	✘	ICCV 2019
DepthTrack [48]	✔	✔	✘	200	294.5K	90	✔	ICCV 2021
ARKitTrack [52]	✔	✔	✘	455	229.7K	144	✔	CVPR 2023
RGBDT500	✔	✔	✔	500	203.7K	66	✔	NeurIPS 2025

APFNet [44] introduces an adaptive feature fusion mechanism that fine-tunes attributes between different modalities, thus enhancing the robustness and accuracy of the algorithm. Furthermore, with the advent of Vision Transformers [8], RGB-T tracking technology has seen further innovations. For example, by introducing visual prompt learning [53, 42, 40, 12] into the Transformer-based tracking frameworks, auxiliary thermal information is fused into pre-trained RGB-only models, significantly improving tracking performance. Additionally, BAT [3] and TBSI [15] propose feature bridging mechanisms to strengthen cross-modal interactions.

It is also noteworthy that research on pixel-level RGB-T image fusion provides valuable guidance for cross-modal information integration in tracking [25]. The comprehensive review by [28], which systematically explores the field from data compatibility to task adaptation, offers a solid theoretical foundation for understanding how to effectively fuse pixel-level heterogeneous modalities to serve high-level vision tasks [17, 27]. For instance, CoCoNet [26] introduces a multi-level feature ensemble for multi-modal image fusion, significantly advancing the performance in both image fusion and downstream object detection tasks. Through pixel-level fusion, complementary information from RGB and TIR modalities can be effectively aligned and enhanced [35], thereby providing a more robust and information-rich joint representation for subsequent trackers.

However, most current multi-modal tracking methods are predominantly designed for dual-modality scenarios, typically combining RGB with either depth or thermal information. As such, they are not well-suited to directly process and integrate tri-modal data due to architectural limitations and the distinct characteristics of each modality. In contrast, our proposed RDTTrack baseline is capable of jointly leveraging RGB, depth, and thermal inputs, thus improving tracking robustness. In addition, its architecture is tailored to handle the heterogeneity across modalities.

3 The RGBDT500 Dataset

To promote the development of more general multi-modal object tracking and to support tri-modal tracking task, this work introduces the RGBDT500 dataset. The RGBDT500 contains three modalities, including RGB, depth and thermal infrared, encompassing a wide range of object categories and challenging scenarios for tracking.

3.1 Dataset Details

The RGBDT500 dataset comprises 400 training sequences, and 100 test sequences, with approximately 160K tri-modal frames in the training set and 43.7K tri-modal frames in the test set. For each tri-modal frame, the RGB and depth images are captured using the ZED stereo camera, while the thermal infrared image is recorded using a separate LGCS121 thermal camera. The ZED stereo camera delivers time-synchronised and pixel-aligned RGB and depth image pairs. To resolve the resolution mismatch between the thermal infrared images and the RGB/depth modalities, we employ a manually feature point matrix mapping approach that aligns the target region and surrounding pixels across all three modalities. Furthermore, all tri-modal images are stored in PNG format with a uniform resolution of 1920×1080.

The RGBDT500 dataset includes a diverse range of object categories, covering over 66 classes, including household items, animals, and vehicles, as shown in Fig. 2. In constructing the categories, we strategically consider the unique strengths and limitations of each modality. For example, sequences featuring objects in low-light conditions or containing high-temperature objects are deliberately captured to emphasise the advantages of the thermal modality. Conversely, we capture several video sequences featuring planar objects with minimal geometric depth variation to intentionally limit the contribution of the depth modality, thereby encouraging greater reliance on RGB and thermal infrared cues. To further enhance the dataset's complexity and realism, we capture some sequences exhibiting modal perturbations, such as inconsistent viewpoints among RGB, depth, and thermal infrared streams. Collectively, these design choices make RGBDT500 a challenging and comprehensive benchmark for evaluating the robustness of tri-modal tracking algorithms. In addition, to preserve privacy, we have applied EgoBlur [33] model to blur all recognizable faces and license plates present in the images.

Table 1 presents a comparison between the proposed RGBDT500 dataset and related visual tracking datasets. As shown, compared to the current multi-modal tracking dataset, the RGBDT500 dataset offers several key advantages. It provides spatially and temporally aligned RGB, depth and TIR modalities for tri-modal tracking, which existing datasets limited to RGB-D and RGB-T pairs cannot support. Besides, compared to most RGB-D and RGB-T tracking datasets, RGBDT500 offers a clear advantage in the number of sequences, total frames and object classes.

3.2 Data Annotation

The RGBDT500 is annotated with accurate object bounding boxes for frames containing the object of interest, providing reliable ground truth for developing and assessing tracking algorithms. For the training set, to minimise annotation costs and while meeting requirements for training, we employ K-means clustering [10] to select the most representative frames from each training sequence, which are subsequently annotated with object bounding boxes. Besides, the test set includes 100 sequences, encompassing 43.7K RGB-D-T image triplets with dense bounding box annotations, thereby facilitating rigorous and detailed performance assessment. For the bounding box annotation, we adopt the top-left corner coordinates (x, y), along with the width w and height h of the target's bounding box, to represent the ground truth in the format $[x, y, w, h]$.

3.3 Evaluation Metrics

For evaluation, we follow the One-Pass Evaluation (OPE) protocol [41] to assess tracking performance on RGBDT500. Specifically, we compute the Distance Precision (DP) and the Area Under the Curve (AUC) of the success plot to quantitatively measure the effectiveness of multi-modal tracking methods. Specifically, the precision is measured by computing the distance between the predicted and the ground-truth bounding boxes. By varying a predefined distance threshold to determine successful tracking, a precision plot can be generated to illustrate tracker performance. Based on the precision

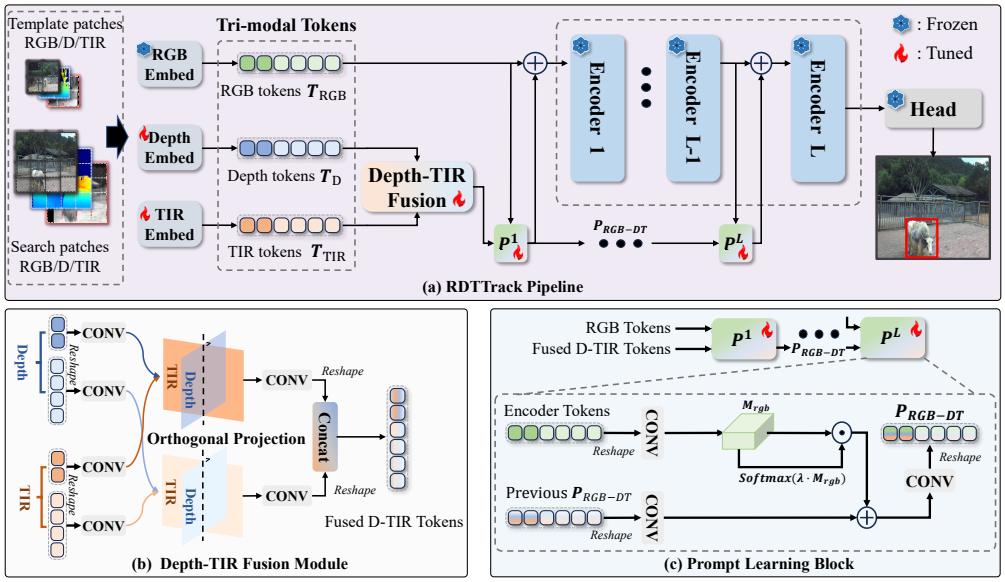


Figure 3: An overview of the pipeline and architecture of RDTTrack. (a) illustrates the overall tracking pipeline; (b) depicts the detailed structure of the Depth-TIR fusion module; and (c) presents the architecture of the prompt learning block.

plot, DP is defined as the percentage of frames in which the prediction error falls within a threshold of 20 pixels. In addition, tracking success can also be evaluated by determining whether the overlap ratio between the predicted and ground-truth bounding boxes exceeds a predefined threshold. The success plot illustrates the proportion of successful frames across a range of overlap thresholds from 0 to 1. Then, the AUC of the success plot is computed to quantitatively assess trackers.

4 The Tri-Modal Tracking Baseline

To support tri-modal object tracking on RGBDT500 and promote progress toward more general multi-modal tracking, we design a specialised tracking algorithm, RDTTrack, capable of effectively integrating RGB, depth, and thermal infrared information. The RDTTrack is developed by extending a pre-trained RGB-only tracker, OSTrack [49], with a specially designed prompt learning module that enables integration of tri-modal cues for tracking.

4.1 Overall Architecture

The overall pipeline and architecture of RDTTrack are illustrated in Fig. 3. The pre-trained OSTrack model consists of a standard patch embedding module, L Transformer encoder layers and a bounding box prediction head. As shown in Fig. 3(a), firstly, the tri-modal template patches ($\boldsymbol{Z}_M \in \mathbb{R}^{3\times H_z\times W_z}$, where $M \in \{\mathrm{RGB}, \mathrm{D}, \mathrm{TIR}\}$) and search patches ($\boldsymbol{X}_M \in \mathbb{R}^{3\times H_x\times W_x}$) are first processed through a patch embedding layer with positional encoding to produce RGB, depth and TIR template tokens $\boldsymbol{T}_M^Z \in \mathbb{R}^{C\times h_z h_z}$ and search tokens $\boldsymbol{T}_M^X \in \mathbb{R}^{C\times h_x h_x}$ respectively. It is worth noting that the single-channel depth and TIR images are converted into three-channel representations similar to RGB images. Then the corresponding template and search tokens for each modality are concatenated to construct the input tokens $\boldsymbol{T}_M \in \mathbb{R}^{C\times(h_z h_z + h_x h_x)}$ of each modality, calculated as:

$$\mathbf{T}_M = [(\mathrm{PE}(\boldsymbol{Z}_M) + \mathrm{Pos_Z}) \Vert_2 (\mathrm{PE}(\boldsymbol{X}_M) + \mathrm{Pos_X})], M \in \{\mathrm{RGB}, \mathrm{D}, \mathrm{TIR}\}, \quad (1)$$

where PE is the patch embedding operation, and $\mathrm{Pos_Z}$ and $\mathrm{Pos_X}$ are positional encodings. $[\cdot\Vert_2\cdot]$ denotes the concatenation operator along the token dimension (the second dimension).

Subsequently, the depth and thermal infrared tokens are integrated as auxiliary multi-modal information to enhance the pre-trained RGB-only tracking OSTrack model. Specifically, the RGB tokens, together with the fused depth and thermal infrared (D-TIR) tokens, are fed into a prompt learning

block designed to learn effective multi-modal visual prompts. Then, the multi-modal visual prompts are added with the RGB tokens and input to the L-layer pre-trained vision Transformer encoder for feature extraction and interaction. The learned multi-modal visual prompts are subsequently added to the RGB tokens and fed into the L-layer pre-trained vision Transformer encoder for feature extraction and interaction. Each encoder layer comprises a multi-head self-attention mechanism, layer normalisation, a feed-forward network, and skip connections. At each encoder layer, the input is constructed by integrating the tokens from the prompt module with the RGB tokens, thereby enabling robust and comprehensive tri-modal feature interaction. The output of the final encoder layer is passed to the prediction head, which produces the tracking results.

For the training of RDTTrack, we freeze all parameters of the RGB streams, and fine-tune only the Depth-TIR fusion module and all prompt learning blocks. The overall loss function of RDTTrack is a combination of focal loss [23] for classification, L_1 loss and the GIoU loss [34] for localisation, calculated as: $\mathcal{L} = L_{\text{CLS}} + \lambda_{\text{GIoU}} \cdot L_{\text{GIoU}} + \lambda_{L_1} \cdot L_1$, where λ_{GIoU} and λ_{L_1} are two constants.

4.2 Depth-TIR Fusion Module

To fully exploit the tracking potential of the auxiliary depth and thermal infrared modalities, we propose a depth-TIR fusion module based on orthogonal projection constraints, as shown in Fig. 3(b). In particular, the depth tokens and TIR tokens are divided into template and search tokens. These tokens are then reshaped into 2D spatial feature maps for further processing. To obtain compact and discriminative representations for each modality, two 1×1 convolutional layers are independently applied to the depth and TIR inputs. Then, depth and TIR inputs are processed through an orthogonal projection, calculated as:

$$\begin{cases} \boldsymbol{F}_D^Z = \boldsymbol{F}_D^Z - \alpha \cdot \dfrac{(\boldsymbol{F}_D^Z \cdot \boldsymbol{F}_{TIR}^Z)}{\|\boldsymbol{F}_{TIR}^Z\| + \epsilon} \cdot \boldsymbol{F}_{TIR}^Z, & \boldsymbol{F}_{TIR}^Z = \boldsymbol{F}_{TIR}^Z - \beta \cdot \dfrac{(\boldsymbol{F}_{TIR}^Z \cdot \boldsymbol{F}_D^Z)}{\|\boldsymbol{F}_D^Z\| + \epsilon} \cdot \boldsymbol{F}_D^Z, \\ \boldsymbol{F}_D^X = \boldsymbol{F}_D^X - \alpha \cdot \dfrac{(\boldsymbol{F}_D^X \cdot \boldsymbol{F}_{TIR}^X)}{\|\boldsymbol{F}_{TIR}^X\| + \epsilon} \cdot \boldsymbol{F}_{TIR}^X, & \boldsymbol{F}_{TIR}^X = \boldsymbol{F}_{TIR}^X - \beta \cdot \dfrac{(\boldsymbol{F}_{TIR}^X \cdot \boldsymbol{F}_D^X)}{\|\boldsymbol{F}_D^X\| + \epsilon} \cdot \boldsymbol{F}_D^X, \end{cases} \tag{2}$$

where the $(.,.)$ is inner product operation. $\boldsymbol{F}_D^Z \in \mathbb{R}^{C \times h_Z \times h_Z}$, $\boldsymbol{F}_D^X \in \mathbb{R}^{C \times h_X \times h_X}$, $\boldsymbol{F}_{TIR}^Z \in \mathbb{R}^{C \times h_Z \times h_Z}$, $\boldsymbol{F}_{TIR}^X \in \mathbb{R}^{C \times h_X \times h_X}$ are the template and search region feature maps of depth and TIR modalities, respectively. α and β are two learnable constant number. ϵ is a very small fixed constant. Through the above process, the orthogonal features of the depth and TIR modalities are effectively extracted, enabling them to complement each other and provide richer, more discriminative information for robust tracking.

Subsequently, $\boldsymbol{F}_D^Z$ and $\boldsymbol{F}_{TIR}^Z$, as well as $\boldsymbol{F}_D^X$ and $\boldsymbol{F}_{TIR}^X$, are concatenated along the channel dimension, respectively. Afterwards, a 1×1 convolution is applied to each concatenated feature map, which is then reshaped back into tokens. The resulting dual-modal template tokens and search tokens are finally concatenated along the token sequence dimension to form a unified D-TIR fused representation $\boldsymbol{T}_{\text{D-TIR}} \in \mathbb{R}^{C \times (h_Z h_Z + h_X h_X)}$. These operations are computed as:

$$\begin{cases} \boldsymbol{F}_{D-TIR}^Z = \text{Conv}([\boldsymbol{F}_D^Z \underset{1}{\|} \boldsymbol{F}_{TIR}^Z]), \quad \boldsymbol{F}_{D-TIR}^X = \text{Conv}([\boldsymbol{F}_D^X \underset{1}{\|} \boldsymbol{F}_{TIR}^X]), \\ \boldsymbol{T}_{D-TIR} = [\text{Reshape}(\boldsymbol{F}_{D-TIR}^Z) \underset{2}{\|} \text{Reshape}(\boldsymbol{F}_{D-TIR}^X)], \end{cases} \tag{3}$$

where $[\cdot \underset{1}{\|} \cdot]$ denotes the concatenation operator along the channel dimension (the first dimension).

4.3 Multi-Modal Prompt Learning Blocks

Fig. 3(c) presents the architecture of the multi-modal prompt learning module. It is designed to extract complementary information between the RGB modality and the fused auxiliary D-TIR modalities to produce more informative visual prompts. The prompt learning block follows the design proposed in [53]. The l-th prompt learning block $\text{Propmt}^l()$ takes as input the tokens $\boldsymbol{H}^{l-1}$ from $(l-1)$-th Transformer encoder layer along with the multi-modal prompts $\boldsymbol{P}^{l-1}$ produced by the preceding prompt block, as:

$$\boldsymbol{P}^l = \text{Propmt}^l\left(\boldsymbol{H}^{l-1}, \boldsymbol{P}^{l-1}\right), \quad l = 1, 2, \ldots, L. \tag{4}$$

Table 2: Comparison of RGDTTrack with multiple SOTA trackers on RGBDT500.

Tracker	Input Modality	AUC	DP	Publication Year
ATOM [6]	RGB	0.649	0.695	2019
DiMP [2]	RGB	0.662	0.710	2019
PrDiMP [7]	RGB	0.676	0.698	2020
ToMP [31]	RGB	0.710	0.747	2022
OSTrack [49]	RGB	0.694	0.736	2022
SeqTrack [4]	RGB	0.719	0.766	2023
MixFormer [5]	RGB	0.732	0.781	2024
DeT_ATOM [48]	RGB+D	0.639	0.665	2021
DeT_DiMP [48]	RGB+D	0.667	0.700	2021
SPT [57]	RGB+D	0.706	0.757	2023
ViPT_RGBD [53]	RGB+D	0.720	0.759	2023
SDSTrack_RGBD [13]	RGB+D	0.718	0.763	2024
UnTrack_RGBD [42]	RGB+D	0.733	0.776	2024
TBSI [15]	RGB+T	0.690	0.749	2023
ViPT_RGBT [53]	RGB+T	0.693	0.752	2023
SDSTrack_RGBT [13]	RGB+T	0.666	0.708	2024
BAT [3]	RGB+T	0.713	0.782	2024
UnTrack_RGBT [42]	RGB+T	0.732	0.790	2024
DCEvo+OSTrack [29]	RGB+T	0.706	0.741	2025
RDTTrack(Ours)	RGB+D+T	**0.752**	**0.792**	2025

Table 3: Tracking efficiency comparison of several trackers.

Tracker	RDTTrack	BAT	TBSI	SDSTrack	Un-Track	ViPT
FPS	**76.6**	26.8	32.5	20.9	41.2	31.8

In the prompt learning block, $\boldsymbol{P}^{l-1}$ and $\boldsymbol{H}^{l-1}$ are projected to a space of reduced channel dimension using a 1×1 convolutional layer. Then the embeddings $\boldsymbol{H}^{l-1}$ are enhanced by a spatial fovea operation, which adopts a λ-smoothed softmax across all the spatial dimensions. Finally, the multi-modal prompt embeddings are generated by adding $\boldsymbol{P}^{l-1}$ to the enhanced $\boldsymbol{H}^{l-1}$ and a 1×1 convolutional layer. The detailed operations are calculated as:

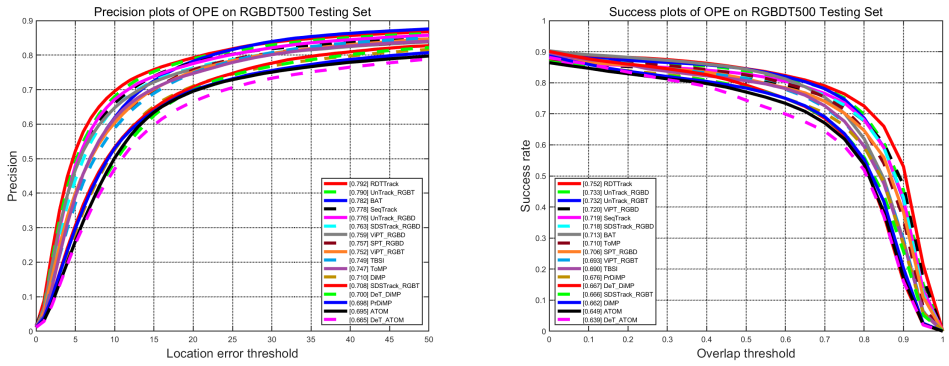
$$\begin{cases} \boldsymbol{A}_{\mathrm{RGB}} = \mathrm{Conv}\left(\boldsymbol{H}^{l-1}\right), \quad \boldsymbol{A}_{P} = \mathrm{Conv}\left(\boldsymbol{P}^{l-1}\right) \\ \boldsymbol{A}^{e}_{\mathrm{RGB}} = \boldsymbol{A}_{\mathrm{RGB}} \odot \boldsymbol{A}_{\mathrm{fovea}}, \quad \boldsymbol{A}_{\mathrm{fovea}} = \left\{ \frac{e^{\boldsymbol{A}_{\mathrm{RGB}}[i,j]}}{\sum e^{\boldsymbol{A}_{\mathrm{RGB}}[i,j]}} \lambda \right\}, \\ \boldsymbol{P}^{l} = \mathrm{Conv}\left(\boldsymbol{A}^{e}_{\mathrm{RGB}} + \boldsymbol{A}_{P}\right) \end{cases} \tag{5}$$

$\odot$ denotes element-wise multiplication. For the first prompt learning block, its input is the RGB tokens $\boldsymbol{T}_{\mathrm{RGB}}$ and the D-TIR fused representation $\boldsymbol{T}_{\mathrm{D-TIR}}$. For more detailed information of the prompt learning block, readers are referred to the reference [53].

5 Experiments

5.1 Experimental Settings

Our proposed RDTTrack is implemented in Python 3.8 with PyTorch 1.12, and all training and evaluation procedures are conducted on a single NVIDIA RTX 3090 GPU. The RDTTrack is fine-tuned on the RGBDT500 dataset for 60 epochs, with each epoch containing 60K tri-modal samples. The frozen parameters of RDTTrack are initialised using the pre-trained weights of the OSTrack model. The initial learning rate is set to $4e-5$ and is decreased by a factor of 10 after the 48-th epoch. All evaluations are performed using the standard OPE tracking protocol, including DP, AUC and tracking speed Frames Per Second (FPS) metrics.



(a) (b)

Figure 4: The precision plots and success plots of trackers on RGBDT500.

5.2 Comparative Experiments

We evaluate a range of RGB-only, RGB-D and RGB-T trackers, inclding ATOM [6], DiMP [2], PrDiMP [7], ToMP [31], OSTrack [49], SeqTrack [4], Mixformer [5], DeT [48], SPT [57], ViPT [53], SDSTrack [13], TBSI [15], BAT [3], UnTrack [42], and DCEvo [29] fusion with OSTrack, and compare them with the proposed RDTTrack. Table 2 reports the tracking performance of various methods on the RGBDT500 test set using AUC and DP metrics. Our RDTTrack achieves 0.752 and 0.792 in AUC and DP, respectively. As shown, the proposed RDTTrack demonstrates superior performance across both metrics, consistently outperforming existing uni-modal and dual-modal trackers. Specifically, among single-modal RGB-only trackers, the MixFormer achieves the best DP and AUC of 0.732 and 0.781. By effectively leveraging additional depth and thermal infrared modalities, RDTTrack achieves superior performance compared to MixFormer, with improvements of 2.0% in AUC and 1.1% in DP, respectively.

For dual-modal tracking approach, while the best RGB-D and RGB-T trackers outperform single-modal approaches, they remain inferior to our tri-modal tracking framework. In detail, ViPT_RGBD, SDSTrack_RGBD and UnTrack_RGBD achieve AUC scores of 0.720, 0.718 and 0.733, and DP scores of 0.759, 0.763 and 0.776, respectively. In comparison, RDTTrack surpasses these two RGB-D tracking approaches with an AUC improvement of 3.2%, 3.4% and 1.9% respectively, and a DP improvement of 3.3%, 2.9% and 1.6%, respectively. Among the RGB-T trackers, the UnTrack_RGBT achieves the optimal AUC of 0.732 and DP of 0.790, outperforming other RGB-T methods like ViPT_RGBT, SDS_RGBT and BAT. In addition, the RGB-T image fusion method DCEvo [29] with OSTrack achieves improved performance than RGB-only OSTrack. It clearly validates the effectiveness of image-level fusion of DCEvo for robust multi-modal tracking. The proposed tri-modal tracker, RDTTrack, which incorporates an additional depth modality compared to BAT, achieves performance gains of 3.9% in AUC and 1.0% in DP, respectively. In summary, these results validate the effectiveness of RDTTrack's architecture in fully exploiting the complementary strengths of RGB, depth, and thermal modalities for enhanced tracking performance.

In Fig. 4, we present the precision plots and success plots of trackers on RGBDT500. The DP and AUC scores are exhibited in the legends of the corresponding plots. From the curves, it is evident that the proposed RDTTrack consistently outperforms other single-modal and dual-modal trackers, demonstrating superior tracking accuracy and robustness across different evaluation metrics. Furthermore, we also provide some qualitative results in the supplementary material to intuitively show the advantages of the proposed RDTTrack.

For efficiency comparison, we conduct runtime evaluations of several trackers using FPS as the metric. All experiments are run on an NVIDIA RTX 3090 GPU, and the results are presented in Table 3. As shown, RDTTrack achieves significantly higher runtime efficiency compared to other recent multi-modal trackers, reaching 76.6 FPS. This achievement is owes to the use of lightweight prompt learning and a frozen baseline, which reduces the number of additional parameters and the

Table 4: Performance comparison of retrained several trackers.

Tracker	Modality of Training Set	AUC	DP
STARK [47]	RGB	0.692	0.732
SPT_RGBD [57]	RGB+Depth	0.706	0.757
SPT_RGBT [57]	RGB+Thermal	0.730	0.784
OSTrack [49]	RGB	0.694	0.736
ViPT_RGBD [53]	RGB+Depth	0.719	0.762
ViPT_RGBT [53]	RGB+Thermal	0.729	0.768
RDTTrack	RGB+Depth	0.737	0.776
RDTTrack	RGB+Thermal	0.734	0.774
RDTTrack	RGB+Depth+Thermal	**0.752**	**0.792**

resulting computational overhead. In addition, the RDTTrack model contains only 0.86M trainable parameters, making it highly efficient for training as well.

5.3 Ablation Studies

In Table 4, we present the performance of SPT and ViPT on the RGBDT500 test set, following their retraining on the RGBDT500 training set. With training on dual-modal RGB-D and RGB-T data of RGBDT500, the SPT and ViPT achieve oblivious performance improvements compared to their single-modal baselines STARK [47] and OSTrack [47]. Nevertheless, our proposed RDTTrack, which integrates RGB, depth, and thermal modalities, consistently outperforms the retrained SPT and ViPT models, demonstrating the advantage of tri-modal fusion.

In addition, to verify the effectiveness of input modalities of RDTTrack, we construct RDTTrack with different input modality combinations. The results are shown in Table 4. As observed, by leveraging tri-modal input data, RDTTrack outperforms its dual-modal counterparts. Furthermore, we conduct an experiment to validate the effectiveness of the proposed Depth-TIR module. The results are exhibited in Table 5. As shown, removing the Depth-TIR Orthogonal Projection (OP) leads to a notable performance degradation, with the AUC of RDTTrack dropping significantly from 0.752 to 0.733. Moreover, when the learnable α and β items in Eqn. (5) are removed, RDTTrack exhibits notable performance drops in both AUC and DP, highlighting their importance in effective feature disentanglement and fusion.

Table 5: The ablation study on the Depth-TIR fusion module.

Tracker	AUC	DP
w/o D-TIR OP	0.733	0.773
w/o α&β	0.739	0.779
RDTTrack	**0.752**	**0.792**

6 Conclusion

In this work, we presented the first tri-modal tracking dataset, RGBDT500, and a tri-modal tracker, RDTTrack. RGBDT500 is meticulously constructed to address the challenges of tri-modal fusion of RGB, depth, and thermal infrared modalities, offering a comprehensive benchmark for advancing multi-modal tracking research. In addition, the proposed RDTTrack effectively leverages the complementary information across all three modalities, achieving state-of-the-art performance on RGBDT500 and validating the benefits of tri-modal integration. By releasing this dataset and baseline tracker, we aim to facilitate the development of more robust and generalizable multi-modal tracking methodologies and encourage further exploration into multi-modal fusion for real-world applications.

Acknowledgments and Disclosure of Funding

This work was funded by the National Natural Science Foundation of China (62576152, 62020106012, 62336004), the Basic Research Program of Jiangsu (BK20251624, BK20250104), the China Postdoctoral Science Foundation (2025M771593), the Wuxi Science and Technology Development Fund Project (K20241025), and the Fundamental Research Funds for the Central Universities (JUSRP202504007, JUSRP202501041).

References

[1] UT Benchmark. A benchmark and simulator for uav tracking. In *ECCV*, 2016.

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, pages 6182–6191, 2019.

[3] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multimodal tracking. In *AAAI*, volume 38, pages 927–935, 2024.

[4] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, pages 14572–14581, 2023.

[5] Yutao Cui, Cheng Jiang, Gangshan Wu, and Limin Wang. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE TPAMI*, 46(06):4129–4146, 2024.

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019.

[7] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, pages 7183–7192, 2020.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, pages 5374–5383, 2019.

[10] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, pages 19079–19091, 2024.

[13] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *CVPR*, pages 26551–26561, 2024.

[14] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 43(5):1562–1577, 2019.

[15] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *CVPR*, pages 13630–13639, 2023.

[16] Sajid Javed, Martin Danelljan, Fahad Shahbaz Khan, Muhammad Haris Khan, Michael Felsberg, and Jiri Matas. Visual object tracking with discriminative filters and siamese networks: A survey and outlook. *IEEE TPAMI*, 45(5):6552–6574, 2023.

[17] Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multispectral image stitching via global-aware quadrature pyramid regression. *IEEE TIP*, 33:4288–4302, 2024.

[18] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *ICCVW*, pages 2262–2270, 2019.

[19] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE TIP*, 25(12):5743–5756, 2016.

[20] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *PR*, 96:106977, 2019.

[21] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE TIP*, 31:392–404, 2021.

[22] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *ACM MM*, pages 1856–1864, 2017.

[23] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[24] Chang Liu, Yongsheng Yuan, Xin Chen, Huchuan Lu, and Dong Wang. Spatial-temporal initialization dilemma: towards realistic visual tracking. *Visual Intelligence*, 2(1):35, 2024.

[25] Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 12:1–14, 2024.

[26] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *IJCV*, 132(5):1748–1775, 2024.

[27] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE TCSVT*, 32(8):5026–5040, 2022.

[28] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE TPAMI*, 2024.

[29] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *CVPR*, pages 2226–2235, 2025.

[30] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan. Cdtb: A color and depth visual object tracking dataset and benchmark. In *ICCV*, pages 10013–10022, 2019.

[31] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, pages 8731–8740, 2022.

[32] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. Dal: A deep depth-aware long-term tracker. In *ICPR*, pages 7825–7832, 2021.

[33] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. Egoblur: Responsible innovation in aria, 2023.

[34] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.

[35] Chengcheng Song, Hui Li, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. Refinefuse: an end-to-end network for multi-scale refinement fusion of multi-modality images. *Visual Intelligence*, 3(1):16, 2025.

[36] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, pages 233–240, 2013.

[37] Zhangyong Tang, Tianyang Xu, Xiao-Jun Wu, Xue-Feng Zhu, Chunyang Cheng, Zhenhua Feng, and Josef Kittler. Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution. *IEEE TIP*, 2025.

[38] Zhangyong Tang, Tianyang Xu, Xiaojun Wu, Xue-Feng Zhu, and Josef Kittler. Generative-based fusion mechanism for multi-modal tracking. In *AAAI*, volume 38, pages 5189–5197, 2024.

[39] Zhangyong Tang, Tianyang Xu, Xue-Feng Zhu, Hui Li, Shaochuan Zhao, Tao Zhou, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. Omni survey for multimodality analysis in visual object tracking. *arXiv preprint arXiv:2508.13000*, 2025.

[40] Hongyu Wang, Xiaotao Liu, Yifan Li, Meng Sun, Dian Yuan, and Jing Liu. Temporal adaptive rgbt tracking with modality prompt. In *AAAI*, volume 38, pages 5436–5444, 2024.

[41] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE TPAMI*, 37(9):1834–1848, 2015.

[42] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *CVPR*, pages 19156–19166, 2024.

[43] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, and Aleš Leonardis. Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE TCYB*, 48(8):2485–2499, 2017.

[44] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *AAAI*, volume 36, pages 2831–2838, 2022.

[45] Tianyang Xu, Zhenhua Feng, Xiao-Jun Wu, and Josef Kittler. Adaptive channel selection for robust visual object tracking with discriminative correlation filters. *IJCV*, 129(5):1359–1375, 2021.

[46] Tianyang Xu, Xue-Feng Zhu, and Xiao-Jun Wu. Learning spatio-temporal discriminative model for affine subspace based visual object tracking. *Visual Intelligence*, 1(1):4, 2023.

[47] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, pages 10448–10457, 2021.

[48] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *ICCV*, pages 10725–10733, 2021.

[49] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pages 341–357, 2022.

[50] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. In *ICCVW*, pages 01–10, 2019.

[51] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, pages 8886–8895, 2022.

[52] Haojie Zhao, Junsong Chen, Lijun Wang, and Huchuan Lu. Arkittrack: a new diverse dataset for tracking using mobile rgb-d data. In *CVPR*, pages 5126–5135, 2023.

[53] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023.

[54] Xue-Feng Zhu, Xiao-Jun Wu, Tianyang Xu, Zhen-Hua Feng, and Josef Kittler. Robust visual object tracking via adaptive attribute-aware discriminative correlation filters. *IEEE TMM*, 24:301–312, 2022.

[55] Xue-Feng Zhu, Tianyang Xu, Sara Atito, Muhammad Awais, Xiao-Jun Wu, Zhenhua Feng, and Josef Kittler. Self-supervised learning for rgb-d object tracking. *PR*, 155:110543, 2024.

[56] Xue-Feng Zhu, Tianyang Xu, Zongtao Liu, Zhangyong Tang, Xiao-Jun Wu, and Josef Kittler. Unimod1k: Towards a more universal large-scale dataset and benchmark for multi-modal learning. *IJCV*, 132(8):2845–2860, 2024.

[57] Xue-Feng Zhu, Tianyang Xu, Zhangyong Tang, Zucheng Wu, Haodong Liu, Xiao Yang, Xiao-Jun Wu, and Josef Kittler. Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking. In *AAAA*, volume 37, pages 3870–3878, 2023.

A Technical Appendices and Supplementary Material

A.1 Data Samples

Some samples from the RGBDT500 dataset are presented in Fig. 5. As shown, the proposed RGBDT500 covers a wide range of scenarios and includes diverse object categories, highlighting its versatility for multi-modal tracking task. In these scenarios, the three modalities provide complementary information, which can contribute to the improvement of tracking performance.

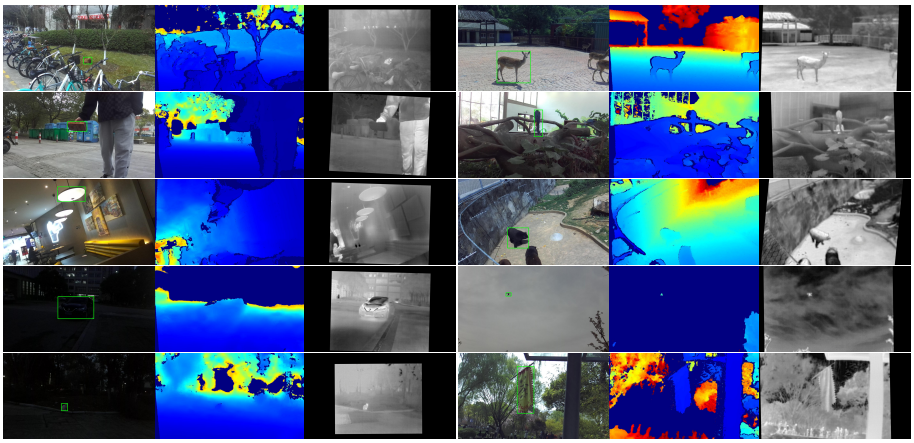


Figure 5: Some samples from the RGBDT500 dataset. The targeted object are highlighted by green bounding boxes.

A.2 Qualitative Analysis

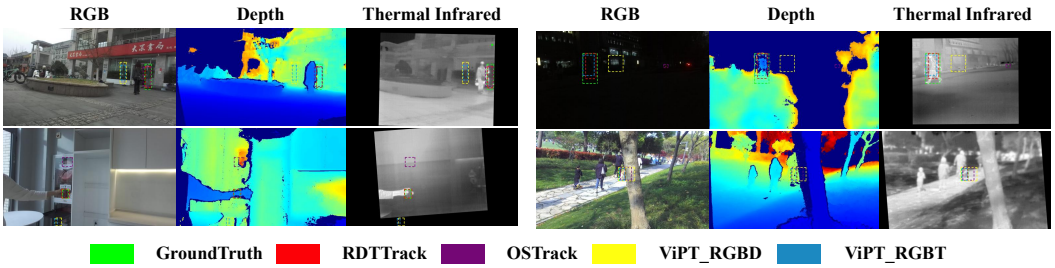


Figure 6: Visualisation of the results of several mainstream trackers on sequences of RGBDT500.

In Fig. 6, we visualized the tracking results of some trackers, including OsTrack (RGB) [49], ViPT_RGBD (RGB + Depth), ViPT_RGBT (RGB + Thermal) [53], and RDTTrack (RGB + Depth + Thermal) on several complex frames of RGBDT500. As shown, in challenging scenarios such as low-light environments or modality interference, our RDTTrack effectively leverages the most reliable modality in each scene to achieve more robust and accurate tracking.

A.3 Limitations and Future Work

Despite the contributions presented in this work, we acknowledge several limitations that pave the way for future research. Specifically, the proposed RDTTrack baseline lacks the flexibility to handle a variable number of input modalities. This restricts its deployment in real-world applications where sensor configurations may vary. In addition, the RGBDT500 dataset, while a significant step forward, could be extended to incorporate other vital modalities, such as LiDAR or event. Future efforts will be directed towards designing a more flexible and input-agnostic fusion architecture, as well as expanding the dataset to include a wider array of sensing modalities.

A.4 Privacy Preservation

The RGBDT500 dataset was collected in controlled environments, with a focus on non-identifiable household objects and animals. In rare scenarios where individuals appear in the dataset, those individuals were members of our research and data collection team, who were fully aware of the recording and gave explicit consent to be included. No bystanders or members of the public were captured without consent. Furthermore, all frames were reviewed to ensure that no identifiable facial features or license plates are present in the released dataset. The potential identifiable facial and license plates are anonymized through blurring by using EgoBlur model.

In addition, all sequences involving public capture occurred in areas where photography is legally permitted. For any such sequences, signage was displayed to indicate that recording was taking place for academic research purposes. Additionally, we have set up a dedicated contact email (xuefeng.zhu@jiangnan.edu.cn) on the dataset website, where individuals can request takedown of specific sequences if they believe they are represented.

A.5 Broader Impacts and Safeguards

This work enhances the robustness of multi-modal tracking, presenting both potential positive and negative societal impacts. On the positive side, it can enable life-saving applications such as improving autonomous vehicle safety in adverse weather conditions and advancing wildlife monitoring efforts. Conversely, the enhanced tracking capabilities across RGB, depth, and thermal domains may be misused in surveillance systems, including mass monitoring, protest tracking, or law enforcement applications that could infringe on privacy and civil liberties. Due to the dual-use nature of this technology, it requires responsible development frameworks and ethical deployment guidelines to mitigate risks such as privacy violations. To mitigate such risks, the RGBDT500 dataset is released under a research-only, non-commercial license that explicitly prohibits use for surveillance, military, or law enforcement purposes. Access to the download link of RGBDT500 requires click-through acceptance of these terms. All sequences involving individuals feature team members with explicit consent, and no identifiable information is present. We encourage transparency and welcome community feedback to support responsible use.

A.6 Dataset Card

- Geographic distribution: Data was collected across distinct regions, including urban, suburban, and indoor lab settings, representing diverse environments and sensor conditions.
- Day/Night ratio: Approximately 80% of the sequences were recorded during the day and 20% at night, to reflect a variety of lighting scenarios.
- Demographic information: As the dataset primarily includes objects and animals in controlled environments, demographic attributes are not labelled or applicable. In rare cases involving people, subjects were team members or individuals who provided consent, and no demographic classification was performed.
- Fairness audit plan: While demographic analysis is not applicable for this version of the dataset, we include a short plan to expand fairness auditing in future iterations. This includes evaluating the dataset's impact in downstream models, e.g., bias in object tracking under varying conditions, and inviting feedback from users to inform gaps in representativeness.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A comprehensive discussion of the limitations of our proposed approach is provided in the supplemental material (see A.3 Limitations and Future Work) to ensure clarity and completeness.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is focused on dataset construction and algorithm design for multi-modal tracking.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of the dataset construction, model architecture, training protocols, and evaluation metrics. Besides, the dataset and codes are provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The RGBDT500 dataset and RDTTrack implementation are made publicly available on the website: https://xuefeng-zhu5.github.io/RGBDT500/.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed information on the training and testing settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provides information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper follows ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion on both potential positive societal impacts and negative societal impacts is presented in supplementary material (see A.5 Broader Impacts and Safeguards).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The risk for misuse of the RGBDT500 dataset and corresponding safeguards are discussed in supplementary material (see A.5 Broader Impacts and Safeguards).

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper makes appropriate references to all existing assets used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces a new tri-modal tracking dataset, RGBDT500, and provides detailed descriptions of its collection, composition, and usage protocols.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The human subjects involved in this work are discussed in supplementary material (see A.4 Privacy Preservation).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The consent and IRB are presented in supplementary material and the data collection section.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.