All-Day Multi-Camera Multi-Target Tracking

Huijie Fan[1,3], Yu Qiao[1], Yihao Zhen[1,3], Tinghui Zhao[1,3], Baojie Fan[4], Qiang Wang[2*]
[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[2]Key Laboratory of Manufacturing Industrial Integrated Automation, Shenyang University
[3]University of Chinese Academy of Sciences [4]Nanjing University of Posts and Telecommunications
{fanhuijie,qiaoyu,zhenyihao,wangqiang}@sia.cn, zhaotinghui22@mails.ucas.ac.cn, jobfbj@gmail.com

Abstract

The capability of tracking objects in low-light environments like nighttime is crucial for numerous real-world applications. However, previous Multi-Camera Multi-Target(MCMT) tracking methods are primarily focused on tracking during daytime with favorable lighting, overlooking the challenge posed by low-light conditions. The main difficulty of tracking under low-light condition is the lack of detailed visible appearance features. To address this issue, we incorporate the infrared modality into MCMT tracking framework to provide more useful information. We constructed the first Multi-modality (RGBT) Multi-camera Multi-target tracking dataset named ***M3Track****, which contains sequences captured in low-light environments, laying a solid foundation for all-day multi-camera tracking. Based on the proposed dataset, we propose All-Day Multi-Camera Multi-Target tracking network, termed as* ***ADMCMT****. Specifically, we propose an All-Day Mamba Fusion(ADMF) module to fuse information from different modalities adaptively. Within ADMF, the Lighting Guidance Model(LGM) extracts lighting relevant information to guide the fusion process. Furthermore, the Nearby Target Collection(NTC) strategy is designed to enhance tracking accuracy by leveraging information derived from surrounding objects of targets. Experiments conducted on M3Track demonstrate that ADMCMT exhibits strong generalization across different lighting conditions. The code will be released at https://github.com/QTRACKY/ADMCMT.*

1. Introduction

Multi-Camera Multi-Target (MCMT)[21, 22] tracking aims to locate and associate the same targets between different frames and camera views with substantial overlaps. Compared with single-camera tracking, using multiple cameras to track the same target can effectively address the occlusion problem and improve tracking consistency. MCMT tracking has significant implications for a number of areas, including urban reconnaissance[20], crowd behavior analysis[28], and traffic scene understanding[32].

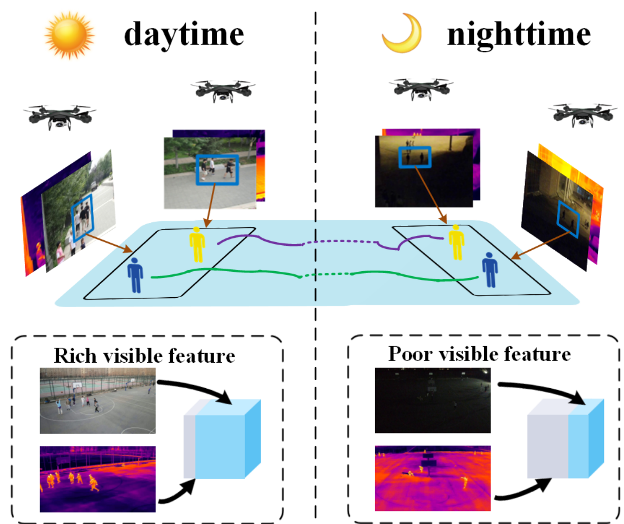

Figure 1. All-day tracking is dedicated to achieving robust tracking at any time of a day. Our approach to achieve all-day MCMT tracking is introducing infrared modality to MCMT framework, employing the complementary nature of infrared and visible light information to empower tracker with capability of tracking objects under different lighting conditions.

Previous MCMT tracking methods[3, 12, 29, 30] focused primarily on tracking at daytime with favorable lighting, overlooking the challenge posed by low-light (nighttime) conditions, which is inevitable in those aforementioned applications. A main limitation that obstructs previous MCMT methods to perform low-light tracking is the lack of detailed visible features. Inspired by multi-modality single-camera tracking methods[2, 14, 34, 48, 50], we incorporate infrared modality into MCMT tracking framework to introduce more detailed information, as the in-

[*]Corresponding author

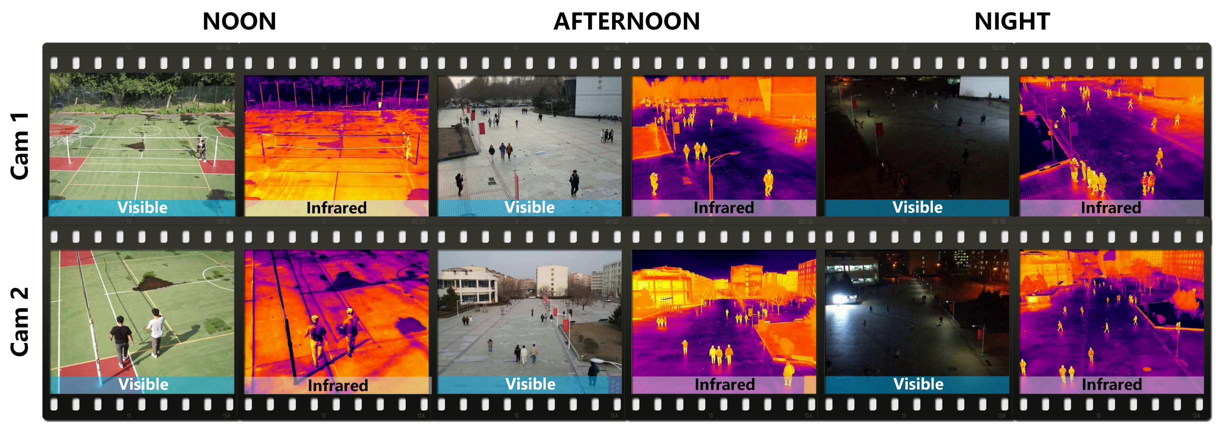

Figure 2. Example of the M3Track. M3Track is composed of well-aligned RGB and infrared sequences captured in different times of a day. From left to right: noon, afternoon, night.

frared modality is hardly affected by lighting conditions. As shown in Fig.1, our core idea is using visible light (RGB) and infrared (Thermal) information as complementary to promote tracking performance under low-light condition, so as to realize all-day tracking.

To address the limitation of existing MCMT datasets[8, 39], we constructed a high-quality Multi-modality (RGBT) Multi-camera Multi-target dataset, named M3Track. As shown in Fig 2, the proposed M3Track consists of well aligned RGB and infrared sequences captured at different times of a day and various weather conditions. To the best of our knowledge, our M3Track is the first MCMT tracking dataset that incorporates both infrared modality and low-light sequences. Furthermore, we present All-Day Multi-camera Multi-target tracking network termed as ADMCMT, which can jointly perform modality fusion, target detection, and target tracking at any time of a day. Specifically, we propose a modality fusion module based on mamba[10], named All-Day Mamba Fusion (ADMF). Within the ADMF, the Lighting Guidance Model (LGM) is used to extract lighting relevant information from visible light input, then the extracted lighting relevant information is sent into mamba-based fusion channels to guide the fusion of visible and infrared features. After that, the fused feature is sent to CenterNet[6] for detection and a transformer-based network for tracking. During the tracking stage, we employ the Nearby Target Collection strategy (NTC) that leverages information from surrounding objects of target to enhance both single-camera and multi-camera tracking capabilities of our model. Experiments demonstrate that ADMCMT exhibits strong generalization capabilities under different lighting conditions.

Our main contributions are summarized as follows:

- We construct the first Multi-modality (RGBT) Multi-camera Multi-target tracking dataset, named M3Track. It contains sequences captured at different times of a day, laying a solid foundation for all-day MCMT tracking.
- We propose an All-Day Multi-Camera Multi-Target tracking network, termed as ADMCMT, which can adaptively fuse features from different modalities.
- We propose an All-Day Mamba Fusion module which use lighting relevant information to guide feature fusion and a Nearby Target Collection strategy to better utilize background information of targets.

2. Related Work

2.1. MCMT Tracking

Multi-camera multi-target [21, 22] tracking aims to locate and associate targets between frames and cameras with substantial overlaps. Current MCMT tracking methods [3, 12, 13, 29, 31, 39, 40] can be broadly classified into two categories. One approach is to calculate the transformation matrix between images captured by different cameras through point-matching methods. The transformation matrix enables the matching of targets across images. MIA-Net [26] employs SIFT to perform point-matching, enhancing the efficiency of tracking small targets with limited visible features. The main limitation of this type of method is its reliance on similarity between the images, which can degrade performance when there are substantial angular differences between camera viewpoints.

Another approach focuses on matching the appearance features of targets. CrossMOT[11] introduces both single-camera Re-ID embeddings and cross-camera Re-ID embeddings to match features between frames and cameras. GMT[7] constructs features with target appearance and contextual information such as location and time. Features are then used to generate global correlation via a transformer-based model in one-stage manner. Both the point-matching method and target matching method rely on visible infor-

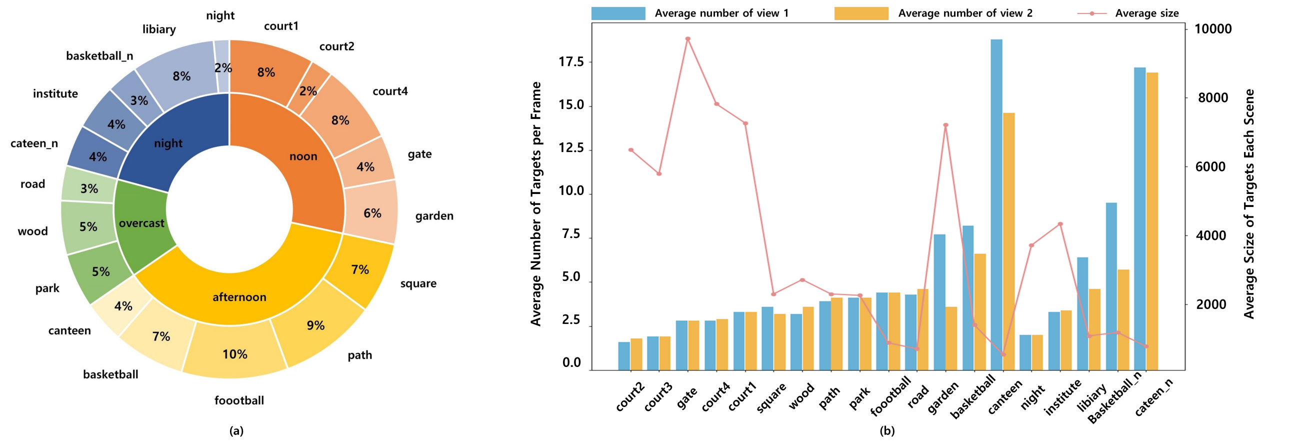

Figure 3. M3Track is collected in diverse scenes under various lighting conditions. (a) represents the proportion of frames for each scene. (b) shows the number and average area of targets each scene, illustrate that M3Track has a high diversity in target dense and size.

Dataset	Scenes	Views	Frames	Boxes	Moving camera	Low-light	UAV view	Multimodality
EPFL	5	4	97K	625K	×	×	×	×
CAMPUS	4	4	83K	490K	×	×	×	×
MvMHAT	1	4	31K	208K	✓	×	×	×
WILDTRACK	1	7	3K	40K	×	×	×	×
DIVOTrack	10	3	54K	560K	✓	×	one	×
M3Track	19	2	118K×2	1188K	✓	✓	two	✓

Table 1. Comparison with previous MCMT datasets. M3Track significantly surpassing previous datasets in both scene diversity and volume.

mation, this dependency may seriously affect the tracking performance under low-light conditions.

2.2. Tracking under low-light condition

Tracking objects in low-light scenes poses a significant challenge owing to the absence of detailed visible features. Approaches to uncover more information in visible light images such as low-light enhancement and domain transfer[5] have been largely explored. LTrack [36] achieves low-light tracking by extracting invariant features from low-light videos utilizing the adaptive low-pass downsample module and the degradation suppression learning strategy. LDEnhancer [42] learns light distribution information to suppress image enhancement and refine image content.

Using extra modalities such as event[45], depth[41], and thermal[35] gives a more direct approach for tracking objects at nighttime by introducing more information. Many multi-modality tracking datasets with low-light sequences have been established[15–19, 47] and played key roles in the generalization of single-target tracking. Methods with low-light tracking ability have also been widely researched on the basis of the rich dataset. BAT[2] uses bi-directional adapter to fuse multi-modality information in an adaptive manner. CMD[48] utilizes a three-stage framework to bridge the performance gap between a compact student RGBT tracking model and a powerful teacher model.

2.3. RGBT Fusion

Infrared and visible light (RGB) data can serve as complementary to provide more useful information, which is essential for a range of downstream tasks, including target detection and tracking. The current methods of modality fusion can be broadly categorized into three types: pixel-level fusion[4], feature-level[2, 24, 27, 38, 46, 51] fusion and decision-level fusion[33]. Pixel-level fusion preserved more detailed information of the input images, but at the cost of high computational complexity. Feature-level fusion fuses features of detected targets, and yields an optimal equilibrium between performance and efficiency. Decision-level fusion processes different modality separately and only fuse the final decision, has high requirements for single modal algorithms. Recently, Mamba has been shown to outperform the transformer in long-term dependency modeling tasks owing to its selective structured state space and has shown competitive results in computer vision tasks. In the field of RGBT fusion, methods based on the Mamba approach[23] are also being actively applied and have shown remarkable performance.

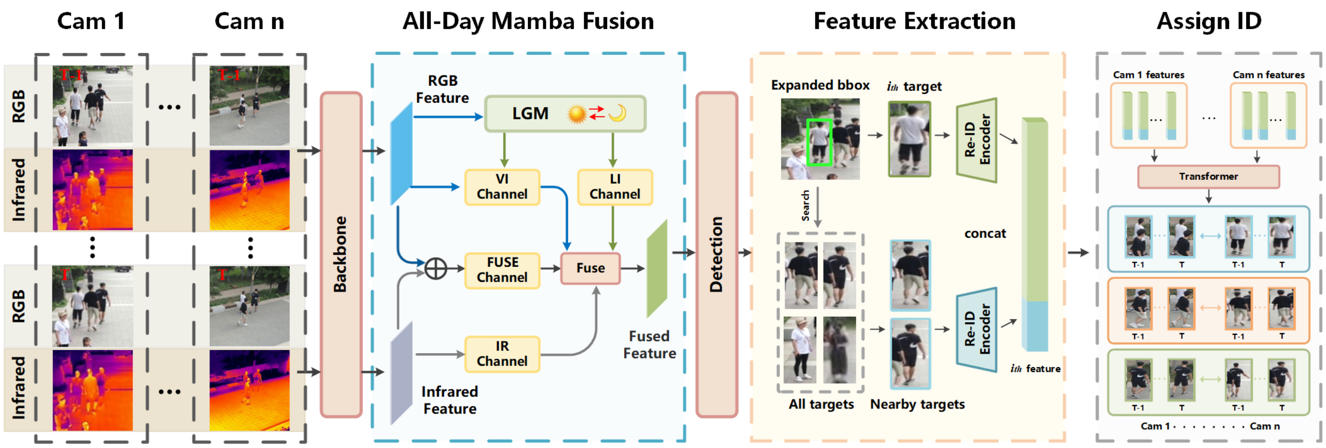

Figure 4. The ADMCMT model first utilizes All-Day Mamba Fusion to fuse infrared and visible light feature under the guidance of lighting relevant information. Fused feature is used for target detection. After obtaining the detection result, Nearby Target Collection is used to help build Re-ID features of targets. Finally, Re-ID features are sent into a transformer based tracking model to obtain the tracking result.

3. M3Track Dataset

3.1. Previous MCMT Datasets

Existing MTMC tracking datasets with overlapping fields of view include: EPFL[8], CAMPUS[39], WILDTRACK[9], MvMHAT[9] and DIVOTrack[11]. Each of these datasets suffers from issues like lack of scene diversity and varying annotation quality. Moreover, all of these datasets are collected at daytime with acceptable lighting, which poses tremendous challenges in conducting MCMT tracking research under low-light conditions.

To address the limitations and enrich the diversity of MCMT datasets, we constructed the first multi-modality MTMC dataset with low-light sequences, named M3Track, which includes aligned RGB and infrared videos collected at different times of a day and in different weather. We hope data with diverse lighting conditions will promote visible light tracking, meanwhile RGB and infrared modalities can serve as complementarity to empower the model with the ability to track under low-light condition, thereby achieving all-day tracking. Comparison with previous MCMT datasets in Table 1 shows that M3Track surpasses previous datasets by nearly two times in terms of frame number, target number, and scenes. It is also noteworthy that previous RGBT tracking datasets are primarily built for single-object tracking, and our M3Track could also serve as a valuable single-camera multi-object RGBT tracking dataset.

3.2. Data Collection

The proposed M3Track dataset was collected with two drones, each equipped with both an RGB camera and an infrared camera. The drones were moving irregularly while recording. We collected 88 sequences across 19 distinct real-world scenarios with diverse surrounding environments and varying target densities, as illustrated in Fig. 3. To obtain diverse lighting conditions, we also collected the data under various weather conditions and different times of a day. These aforementioned efforts ensure that our dataset has sufficient diversity.

After recording, we manually aligned the timestamps and the spatial relationships between the RGB and infrared cameras on each drone. Due to the difference in FOV and resolution between RGB and infrared cameras, we cropped and down-sampled all images to the resolution of 640×512 for better alignment. The dataset is split by approximately 1:1 as the train set and the test set. Previous MCMT tracking datasets usually split one sequence into two for training and testing. However, this approach may result in correlation between the training and testing data, thereby affecting the reliability of test results. To address this issue, the test set of M3Track comprises novel scenes that are not present in the train set, along with scenes that are included in the train set but have distinct targets and varying target densities.

3.3. Data Annotation

M3Track dataset is annotated based on single-modality, we use RGB video as the reference for sequences recorded under preferable lighting conditions, and infrared video for sequences recorded under low-light conditions. The annotation process of M3Track involves two primary steps. First, we labeled all targets from different views at a given time t and assigned a unique ID to each target in the overlapping fields of view of the two cameras. After that, we use the semi-automatic object tracking software DarkLabel to label the targets with the same ID across consecutive frames. The above steps were repeated until all targets in the video are labeled and assigned IDs. The annotation results are carefully checked and adjusted for better accuracy.

4. Method

4.1. Overview

The overall architecture of ADMCMT is illustrated in Fig. 4. The input of ADMCMT consists of visible and infrared image pairs of each view $V_t^n = \{I_t^{ir}, I_t^{vi}\}$ at time t, n represents the n_{th} view. Given the input images, we extract inter-modality features F_t^{ir} and F_t^{vi} and fuse them through All-Day Mamba Fusion module. The fused feature is then used as the input of the target detector. After acquiring the detection results, the Nearby Target Collection strategy is utilized to find surrounding objects for each target. For the i_{th} target, we use different Re-ID decoder to extract target feature F_i and nearby target features F_n. F_i and F_n are concatenated to get the final target feature F_t^i. The set of features in frame t and $t-1$ are sent into a single-stage MCMT tracker to obtain the tracking result.

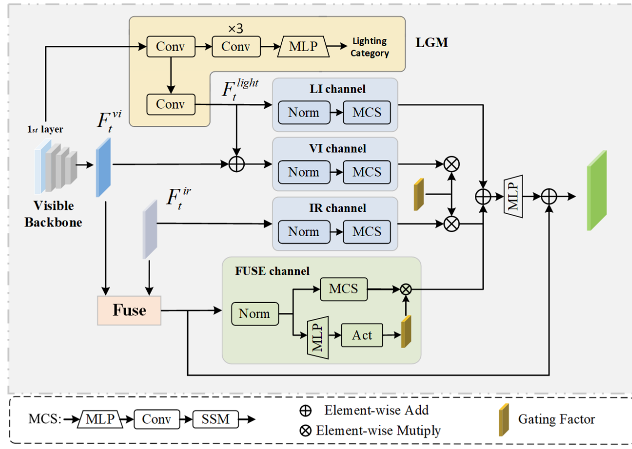

Figure 5. The All-Day Mamba Fusion module consists of the VI, IR, LI channels, and the Lighting Guidance Module(LGM). LGM extracts lighting relevant information F_t^{light} from visible feature F_t^{vi} to guide the fusion process.

4.2. All-Day Mamba Fusion

The richness of information contained in different modalities varies under different lighting conditions. RGB images usually contain more detailed appearance feature than infrared images (e.g color) under preferable lighting condition, while the reverse is true under low-light condition. To more effectively fuse multi-modality information, we proposed **A**ll-**D**ay **M**amba **F**usion (ADMF) model which uses lighting relevant information to guide the fusion progress. The overall architecture of ADMF is illustrated in Fig. 5.

Lighting relevant information. We divided sequences in M3Track into two categories according to their lighting conditions:{Daytime (optimal lighting), Nighttime (poor lighting)}. We send the feature map extracted by the first layer of the visible backbone[44] to the Lighting Guidance Module (LGM), which comprises a stack of four convolution layers and an MLP head to predict the lighting category. The lighting category prediction is constrained by the cross-entropy loss function, which is defined as follows:

$$\mathcal{L}_{\text{light}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right) \quad (1)$$

where $N=2$, y_i is the ground truth for sample i, and $\hat{y}_i$ is the predicted probability. We process the feature extracted by the first convolution layer of LGM with another convolution layer to obtain F_t^{light}. Since the features in LGM are used to predict the lighting category, F_t^{light} should contain information related to lighting condition, thus reflecting the quality of features in RGB image.

Feature fusion. With visible feature F_t^{vi} and infrared feature F_t^{ir} as input, we first fuse them with pixel-level averaging to obtain an initial cross-modality feature F_t^{fuse}. Considering that visible light information is more easily affected by lighting conditions, we add F_t^{light} acquired from LGM to F_t^{vi}. Since F_t^{light} contains information that reflects the quality of features in RGB images, this integration could adjust the weight of information in F_t^{vi} at pixel level. These aforementioned features are further processed by different channels respectively. The VI (F_t^{vi}), IR (F_t^{ir}), and LI (F_t^{light}) channel consist of layer normalization operation and the MCS block, while the FU (F_t^{fuse}) channel is a complete VSS block[25]. The outputs of the VI and IR channels are adjusted by the gating factor derived from the FU channel. Subsequently, all processed features are added and aggregated with an MLP layer. In this process, the lighting relevant information from LI plays a guiding role by indicating the quality of visible information in each pixel. After that, F_t^{fuse} is added to the aggregated feature as residual to obtain the final fusion result.

4.3. Object Detection

We utilize the CenterNet[6] as detector in ADMCMT. CenterNet takes the feature maps derived from ADMF as input and directly predicts the score map of the center of the object, as well as regressing on the size and offset of the object. This approach offers a streamlined and precise solution, making it particularly well-suited to tracking tasks. The location information of the target to be tracked is obtained following the utilization of the CenterNet, which in turn facilitates the extraction of features of the target. The loss function of the target detection model is as follows:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{heatmap}} + \lambda_{\text{size}}\mathcal{L}_{\text{size}} + \lambda_{\text{off}}\mathcal{L}_{\text{off}} \quad (2)$$

where $\lambda_{\text{size}} = 0.1$, $\lambda_{\text{off}} = 1.0$, $\mathcal{L}_{\text{heatmap}}$ denotes the heatmap centroid loss, $\mathcal{L}_{\text{off}}$ denotes the centroid offset loss, and $\mathcal{L}_{\text{size}}$ denotes the target aspect loss.

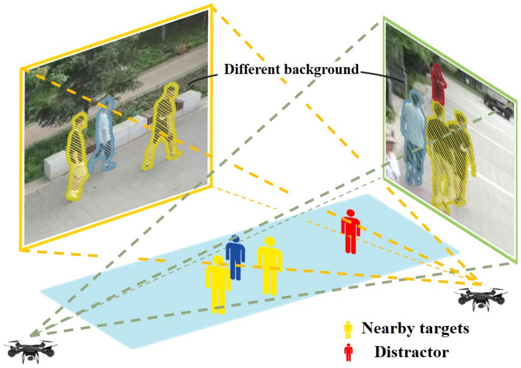

Figure 6. Surrounding objects of target have a high consistency between different views and could be used to improve tracking performance.

4.4. Nearby Targets Collection

Utilizing background information of targets makes a crucial advancement for single-camera tracking[1] by providing the model with the location of the target. However, when it comes to multi-camera tracking, backgrounds of the same target captured by different cameras always have significant difference due to varying shooting angles, as shown in Fig. 6. Following the idea of utilizing background information, we notice that objects near the same target have a high degree of consistency between different cameras. Thus, we develop a **N**earby **T**argets **C**ollection(NTC) strategy to collect information about objects that are close to the target.

After obtaining object detection results in the current frame from all cameras, for the ith target T_i, we expand its bboxes by two times. If there is overlap between the expended bbox and the bbox of another target T_j, we define T_j as a nearby target of T_i. Problem still occurs that objects in one view seem close to each other but are actually far away. To avoid this, we designed a simple area restriction. Specifically, the area of the current target is S, other objects with an area less than $0.6S$ will not be considered as nearby targets. The ith target feature is denoted by F_i and the nearby target are denoted by F_n. Feature dim of each F_n is set to 32 and concatenated with F_i to obtain the final Re-ID feature of target F_t^i.

4.5. Target tracking

We use GMT[7] as tracker in ADMCMT. GMT is a strong transformer based MCMT tracker which can jointly conduct single-camera and cross-camera tracking. GMT takes the set of all target features in two consecutive frames F_t and F_{t-1} as input, output matrix M, where M_{ij} denotes the similarity score between the i_{th} target in F_t and the j_{th} target in F_{t-1}. $M_{i0} = 0$ indicates target i in frame t is not associated with any trajectory in frame t and is therefore designated as $p_i \sim p_0$. For $M_{i0} \neq 0$ and $j > 0$, the highest M_{ij} in row i indicates target i belongs to the trajectory j in frame $t-1$, denoted as $p_i \sim p_j$. The association loss for GMT is as follows:

$$X_{ij} = \begin{cases} 1 & \text{if } p_i \sim p_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\mathcal{L}_{\text{assi}} = -\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}\log(M_{ij}) \tag{4}$$

The loss of ADMCMT is the weighted sum of the losses of the LGM, the target detection module, and the target tracking module.

$$\mathcal{L}_{\text{Track}} = \lambda_{\text{light}}\mathcal{L}_{\text{light}} + \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{assi}}\mathcal{L}_{\text{assi}} \tag{5}$$

where $\lambda_{\text{light}} = 0.1$, $\lambda_{\text{det}} = 2$, and $\lambda_{\text{light}} = 5$. Since $\mathcal{L}_{\text{light}}$ converges most easily , we assign it a minimal weight.

5. Experiments

5.1. Implementation Details

The ADMCMT employs DLA-34 to extract image features and Centernet as the target detector. The model is trained on M3Track with a batch size of 40 (5 frames consist of 2 camera views and 2 modalities) for a total of 20,000 iterations using 4 NVIDIA RTX A6000 GPUs. We utilized Adam optimizer with a learning rate of 10^{-4}. For nearby target collection, the current target is represented by a feature vector of dimension 1024, while each nearby target is represented by a feature vector of dimension 32. The inference is conducted on a single NVIDIA A6000 GPU. The confident threshold of the target detector has been set at 0.525 during inference. Which is the same as the previous works.

5.2. Evaluation Metrics

We evaluate the performance of ADMCMT on both single-camera and multi-camera metrics. We use MOTA, HOTA, IDF1, DetA, and MOTP as evaluation metrics for single-camera tracking and using Multi-Device target Association score (MDA)[26], Cross-View IDF1 (CVIDF1)[11], and Cross-View Matching Accuracy (CVMA) [11] to evaluate cross-camera tracking. MDA evaluates whether a tracker can accurately assign the same ID to the same target across different cameras and is defined as:

$$\begin{aligned} MDA = &\tfrac{1}{C_N^2 * F}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\sum_{i=1}^{F} \\ &\times\left(\tfrac{TA_{(jk,i)}}{GA_{(jk,i)}+FA_{(jk,i)}+MA_{(jk,i)}}\right) \end{aligned} \tag{6}$$

$$C_N^2 = \frac{N!}{(N-2)! \times 2!} \tag{7}$$

	Method	Single camera tracking metrics					Multi-camera tracking metrics		
		MOTA↑	HOTA↑	IDF1↑	DetA↑	MOTP↑	MDA↑	CVIDF1↑	CVMA↑
All	OSNet[49]	68.03	50.33	61.60	55.71	75.78	0.2756	48.50	28.60
	CrossMOT[11]	65.30	44.84	59.91	48.62	75.54	0.2019	51.21	38.51
	CT[37]	68.21	50.38	61.17	55.89	75.80	0.4340	53.65	44.71
	MvMHAT[9]	66.39	48.60	58.96	54.99	75.77	0.2178	46.64	26.72
	AGW[43]	67.62	50.44	62.02	55.54	71.54	0.4235	54.81	45.02
	Ours	**70.94**	**56.62**	**71.47**	**57.89**	**77.21**	**0.6169**	**68.77**	**61.81**
Night	OSNet	66.43	49.95	63.25	54.29	75.86	0.2318	48.39	23.43
	CrossMOT	53.94	42.00	58.62	40.43	71.83	0.1425	49.45	35.84
	CT	66.83	51.04	65.33	54.44	75.88	0.2524	49.25	27.85
	MvMHAT	63.23	48.82	61.03	53.03	75.95	0.2079	46.06	24.36
	AGW	66.43	49.95	63.25	54.29	75.86	0.2318	48.39	23.43
	Ours	**69.71**	**51.19**	**66.02**	**56.40**	**76.50**	**0.4966**	**58.05**	**47.15**

Table 2. Comparison with previous SOTA MCMT trackers on M3Track. All presents results on full dataset and Night result presents results on nighttime scenes. The best results and the second best results are in are shown in red and blue.

V	I	FU	VI+IR	LGM	LI	Single camera tracking metrics					Multi-camera tracking metrics		
						MOTA	HOTA	IDF1	DetA	MOTP	MDA	CVIDF1	CVMA
✓						58.98	51.84	65.25	50.24	76.92	**0.6170**	62.58	53.28
	✓					55.67	51.56	64.84	49.82	76.97	0.5132	61.63	48.01
✓	✓	✓				68.23	53.94	67.87	56.96	76.41	0.5717	63.93	57.22
✓	✓	✓	✓			68.62	54.05	68.40	56.37	76.32	0.5728	64.15	57.34
✓	✓	✓	✓	✓		69.16	54.76	69.06	57.01	77.15	0.5768	66.50	58.48
✓	✓	✓	✓	✓	✓	**70.47**	**56.29**	**70.62**	**57.06**	**77.48**	0.5866	**67.09**	**61.02**

Table 3. Ablation study on different channels and LGM in ADMF. V presents visible light input and I presents infrared input. The best results and the second best results are in are shown in red and blue.

Where i denotes the i-th frame, j and k represent different capture devices. $T_{A(jk,i)}$ and $F_{A(jk,i)}$ represent the number of pairs for which the MCMT tracker correctly and failed to associate a target ID, respectively. $G_{A(jk,i)}$ represents the number of pairs for which the ground truth associates a target ID. $M_{A(jk,i)}$ counts MCMT tracker's false associations that are valid in the ground truth.

CVIDF1 is a cross-view version of IDF1 and is defines as:

$$CVIDF1 = \frac{2 \times CVIDP \times CVIDR}{CVIDP + CVIDR} \tag{8}$$

Where CVIDP and CVIDR represent the Cross-View ID Precision and Cross-View ID Recall, respectively.

CVMA is a cross-view version of MOTA and is defines as:

$$CVMA = 1 - \frac{\sum_t (m_t + fp_t + 2mme_t)}{\sum_t g_t} \tag{9}$$

Where m_t and g_t represent the missed detections at time t and the total true targets from all perspectives at that time, respectively. fp_t and mme_t are the counts of false positives and wrong matching pairs, respectively.

5.3. Comparisons With State-of-The-Art Methods

To demonstrate the effectiveness of our proposed ADMCMT, we conducted experiments on M3Track against other SOTA MCMT trackers. Previous MCMT methods are all designed for single-modality tracking. For a fair comparison, we extract the multi-modality fusion results acquired from our ADMF model and use them as inputs to train other models. Moreover, we shared the detection results of our ADMCMT with other SOTA models. The result metrics are presented in Table. 2 . For single-camera tracking, our model achieved a performance of 70.94 MOTA and 71.47 IDF1, outperforming the second best model by 2.73 MOTA and 9.45 IDF1. For multi-view tracking, our model achieved 61.81 CVMA and 68.77 CVIDF1, reported a huge improvement of 16.79 CVMA, and 13.96 CVIDF1. We also conducted experiment on night scenes in M3Track, our model outperforms the second best model by 2.88

	LGM	LI	Single camera tracking metrics					Multi-camera tracking metrics		
			MOTA↑	HOTA↑	IDF1↑	DetA↑	MOTP↑	MDA↑	CVIDF1↑	CVMA↑
Day			69.55	55.47	69.78	57.11	76.54	0.5848	67.62	63.00
	✓		70.38	56.51	70.79	57.99	77.37	0.5934	68.34	63.82
	✓	✓	71.75	58.44	72.92	59.06	77.82	0.6342	70.63	66.59
Night			66.01	50.03	64.61	54.55	76.58	0.3930	54.30	43.01
	✓		66.46	50.21	64.84	54.50	75.76	0.4362	57.27	47.27
	✓	✓	67.16	50.43	64.70	55.14	76.59	0.4436	58.52	47.04

Table 4. Ablation study on and LGM and LI channel. Day presents results on daytime sequences and N presents results on nighttime scenes. The best results are shown in red.

Num	MOTA↑	IDF1↑	CVIDF1↑	CVMA↑
0	70.47	70.62	67.09	61.02
1	70.65	70.67	67.86	61.16
2	69.82	69.63	66.22	59.32
3	69.23	69.31	51.46	57.17

Table 5. Ablation study on number of nearby targets. ADMCMT achieves best performance with 1 nearby target.

Area(S)	MOTA↑	IDF1↑	CVIDF1↑	CVMA↑
0.4	70.63	71.28	68.74	61.63
0.5	70.61	70.77	68.21	61.52
0.6	70.94	71.47	68.77	61.81
0.7	70.51	71.26	67.91	60.73
0.8	70.71	71.01	68.01	61.38

Table 6. Ablation study on area threshold. ADMCMT achieves best performance with area threshold of 0.6.

MOTA,0.69 IDF1, 8.6 CVIDF1 and 11.31 CVMA, exhibiting strong generalization across different lighting conditions.

5.4. Ablation Study

All-Day Mamba Fusion. The Ablation results of ADMF are shown in Table. 3. We first use single visible light and infrared modality as input to construct baseline. To verify the effectiveness of the proposed ADMF, we fuse RGB and infrared modality in a simple point-wise average manner and process the fused feature with FU, observe MOTA and DetA improvement and MDA, MOTP deterioration. The reason for this result may be that the simple fusion method introduces useless information. After adding IR and VI channels to the network, we obtain better performance than the simple point average fuse. The introduction of LGM led to comprehensive improvement. Overall results prove the effectiveness of ADMF and its internal units.

We also conduct a detailed analysis of the experimental results under different lighting conditions. Results in Table. 4 indicates adding LGM and LI channels improves the tracking performance under low-light conditions (night) while maintaining reliability during the day, demonstrating that the LGM and LI channels meet our expectations to adaptively fuse information from different modalities.

Nearby Target Collection. We conduct experiments to explore the effect of nearby target number and area threshold in NTC. Ablation results of nearby target number are shown in Table. 5 tells that ADMCMT achieve the best performance on M3Track with 1 nearby targets. We also conduct ablation on area threshold with one nearby target, results are shown in Table. 6. We consider the best targets number and area threshold depending on the camera pitching angle and object density. Specifically, scenes with dense objects and small camera pitching angle require more targets to contain useful information and lower thresholds to avoid distractors.

6. Conclusion

In this work, we implement all-day MCMT tracking by introducing the infrared modality. We have constructed the first Multi-modality (RGBT) MCMT tracking dataset, named M3Track, laying a solid foundation for all-day MCMT tracking. Based on the M3Track dataset, we present the All-Day Multi-Camera Multi-Target tracking network, termed as ADMCMT. We designed an All-Day Mamba Fusion model to adaptively fuse information from different modalities under the guidance of lighting-relevant information. Furthermore, we introduce the Nearby Target Collection strategy, which promotes tracking performance by effectively utilizing background information. Experiments demonstrate that ADMCMT exhibits strong generalization capabilities across different lighting conditions.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (62273339, U24A201397), the LiaoNing Revitalization Talents Program (XLYC2403128) and the fundamental research project of SIA.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *ECCV*, pages 205–221, 2020. 6

[2] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multimodal tracking. In *AAAI*, pages 927–935, 2024. 1, 3

[3] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *CVPR*, pages 5030–5039, 2018. 1, 2

[4] Nedeljko Cvejic, Stavri G Nikolov, Henry D Knowles, Artur Loza, Alin Achim, David R Bull, and Cedric Nishan Canagarajah. The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In *CVPR*, pages 1–7, 2007. 3

[5] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *CVPR*, pages 12666–12676, 2024. 3

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 2, 5

[7] Huijie Fan, Tinghui Zhao, Qiang Wang, Baojie Fan, Yandong Tang, and LianQing Liu. Gmt: A robust global association model for multi-target multi-camera tracking. *arXiv preprint arXiv:2407.01007*, 2024. 2, 6

[8] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE PAMI*, 30(2):267–282, 2007. 2, 4

[9] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *ACMMM*, 2021. 4, 7

[10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[11] Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *IJCV*, 132(4):1075–1090, 2024. 2, 4, 6, 7

[12] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE TIP*, 29:5191–5205, 2020. 1, 2

[13] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *CVPR*, pages 3650–3657, 2013. 2

[14] 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. 1

[15] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *ICCVW*, pages 0–0, 2019. 3

[16] 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.

[17] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *ACMMM*, 2017.

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

[19] 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. 3

[20] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, pages 4610–4617, 2020. 1

[21] Peng Li, Jiabin Zhang, Zheng Zhu, Yanwei Li, Lu Jiang, and Guan Huang. State-aware re-identification feature for multi-target multi-camera tracking. In *CVPR*, pages 0–0, 2019. 1, 2

[22] Yu-Jhe Li, Xinshuo Weng, Yan Xu, and Kris M Kitani. Visio-temporal attention for multi-camera multi-target association. In *ICCV*, pages 9834–9844, 2021. 1, 2

[23] Zhe Li, Haiwei Pan, Kejia Zhang, Yuhua Wang, and Fengming Yu. Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion. *arXiv preprint arXiv:2404.08406*, 2024. 3

[24] Lei Liu, Chenglong Li, Yun Xiao, Rui Ruan, and Minghao Fan. Rgbt tracking via challenge-based appearance disentanglement and interaction. *IEEE TIP*, 2024. 3

[25] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 5

[26] Zhihao Liu, Yuanyuan Shang, Timing Li, Guanlin Chen, Yu Wang, Qinghua Hu, and Pengfei Zhu. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *IEEE TMM*, 25:1462–1476, 2023. 2, 6

[27] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *ICCVW*, pages 0–0, 2019. 3

[28] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019. 1

[29] Duy MH Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *CVPR*, pages 8866–8875, 2022. 1, 2

[30] Jonah Ong, Ba-Tuong Vo, Ba-Ngu Vo, Du Yong Kim, and Sven Nordholm. A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. *IEEE PAMI*, 44(5): 2246–2263, 2020. 1

[31] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip:

A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *CVPR*, pages 13784–13793, 2021. 2

[32] Kyujin Shim, Sungjoon Yoon, Kangwook Ko, and Changick Kim. Multi-target multi-camera vehicle tracking for city-scale traffic management. In *CVPR*, pages 4193–4200, 2021. 1

[33] Zhangyong Tang, Tianyang Xu, and Xiao-Jun Wu. Temporal aggregation for adaptive rgbt tracking. *arXiv preprint arXiv:2201.08949*, 2022. 3

[34] Zhangyong Tang, Tianyang Xu, Xiaojun Wu, Xue-Feng Zhu, and Josef Kittler. Generative-based fusion mechanism for multi-modal tracking. In *AAAI*, pages 5189–5197, 2024. 1

[35] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. In *CVPR*, pages 7064–7073, 2020. 3

[36] Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, and Ying Fu. Multi-object tracking in the dark. In *CVPR*, pages 382–392, 2024. 3

[37] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 212–223, 2021. 7

[38] Qin Xu, Yiming Mei, Jinpei Liu, and Chenglong Li. Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE TMM*, 24:567–580, 2021. 3

[39] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *CVPR*, pages 4256–4265, 2016. 2, 4

[40] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Song-Chun Zhu. Cross-view people tracking by scene-centered spatio-temporal parsing. In *AAAI*, 2017. 2

[41] Jinyu Yang, Shang Gao, Zhe Li, Feng Zheng, and Aleš Leonardis. Resource-efficient rgbd aerial tracking. In *CVPR*, pages 13374–13383, 2023. 3

[42] Junjie Ye, Changhong Fu, Ziang Cao, Shan An, Guangze Zheng, and Bowen Li. Tracker meets night: A transformer enhancer for uav tracking. *IEEE Robotics and Automation Letters*, 7(2):3866–3873, 2022. 3

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE PAMI*, 44(6): 2872–2893, 2021. 7

[44] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, pages 2403–2412, 2018. 5

[45] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, pages 13043–13052, 2021. 3

[46] 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 0–0, 2019. 3

[47] 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. 3

[48] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *CVPR*, pages 5404–5413, 2023. 1, 3

[49] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 7

[50] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023. 1

[51] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust rgbt tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2020. 3