Computational Visual Media
https://doi.org/10.1007/s41095-023-0345-5 Vol. 10, No. 2, April 2024, 193–214

Review Article

Multi-modal visual tracking: Review and experimental comparison

Pengyu Zhang[1], **Dong Wang**[1] (✉), **and Huchuan Lu**[1]

Abstract Visual object tracking has been drawing increasing attention in recent years, as a fundamental task in computer vision. To extend the range of tracking applications, researchers have been introducing information from multiple modalities to handle specific scenes, with promising research prospects for emerging methods and benchmarks. To provide a thorough review of multi-modal tracking, different aspects of multi-modal tracking algorithms are summarized under a unified taxonomy, with specific focus on visible-depth (RGB-D) and visible-thermal (RGB-T) tracking. Subsequently, a detailed description of the related benchmarks and challenges is provided. Extensive experiments were conducted to analyze the effectiveness of trackers on five datasets: PTB, VOT19-RGBD, GTOT, RGBT234, and VOT19-RGBT. Finally, various future directions, including model design and dataset construction, are discussed from different perspectives for further research.

Keywords visual tracking; object tracking; multi-modal fusion; RGB-T tracking; RGB-D tracking

1 Introduction

Visual object tracking is a fundamental task in computer vision, which is widely applied in many areas, such as smart surveillance, autonomous driving, and human–computer interaction. Traditional tracking methods are mainly based on visible (RGB) images captured by a monocular camera. For targets suffering from long-term occlusion or in low-illumination scenes, the RGB tracker does not work well and may cause tracking failure. With the easy-access binocular camera, tracking using multi-modal information such as visible-depth, -thermal, -radar, and -laser, is a prospective research direction that has become popular in recent years. Many datasets and challenges have been presented [1–6]. Motivated by these developments, trackers with multi-modal cues have been proposed with satisfying accuracy and robustness against extreme tracking scenarios [7–11].

However, after the emergence of multi-modal trackers, a comprehensive and in-depth survey has not been conducted. To this end, we revisit existing methods under a unified view and evaluate them on well-known datasets. The contributions of this work can be summarized as follows.

- A substantial review is provided for multi-modal tracking methods from various aspects under a unified view. We exploit the similarity of RGB-D and RGB-T tracking and classify them in a unified framework. The existing 61 multi-modal tracking methods are categorized based on auxiliary modality, tracking framework, and related datasets with corresponding metrics. A taxonomy with detailed analysis covers the main knowledge in this field and provides an in-depth introduction to multi-modal tracking models.
- A comprehensive and fair evaluation of popular trackers is conducted on several datasets. We evaluated 34 methods consisting of 15 RGB-D and 19 RGB-T trackers on five datasets, in terms of accuracy and speed, for various applications. The advantages and drawbacks of different frameworks were further analyzed in qualitative and quantitative experiments.
- A prospective discussion for multi-modal tracking is provided. The potential direction of multi-modal tracking in model design and dataset construction is presented, to provide prospective guidance to researchers.

1 Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China. E-mail: P. Zhang, pyzhang@mail.dlut.edu.cn; D. Wang, wdice@dlut.edu.cn (✉); H. Lu, lhchuan@dlut.edu.cn.

Manuscript received: 2023-01-09; accepted: 2023-03-25

The rest of the paper is organized as follows. In Section 2, we introduce existing related basic concepts and previous related surveys. Section 3 provides a taxonomical review of multi-modal tracking. In Section 4, following an introduction on existing datasets, challenges and corresponding evaluation metrics are described. In Section 5, the experimental results on several datasets and different challenges are reported. Finally, we discuss the future direction of multi-modal tracking in Section 6. All of the collected materials and analysis will be released at `https://github.com/zhang-pengyu/Multimodal_tracking_survey`.

2 Background

2.1 Visual object tracking

Visual object tracking aims to estimate the coordinates and scales of a specific target throughout a given video. In general, tracking methods can be divided into two types based on the information utilized: (1) single-modal tracking and (2) multi-modal tracking. Single-modal tracking locates the target captured by a single sensor, such as laser, visible and infrared cameras, to name a few. In recent years, the computational efficiency, ease of accessibility, and high quality of RGB image tracking, have made it increasingly popular, whereby numerous methods have been proposed to improve tracking accuracy and speed.

In RGB tracking, several frameworks, including the Kalman filter (KF) [12, 13], particle filter (PF) [14, 15], sparse learning (SL) [16, 17], correlation filter (CF) [18, 19], and convolutional neural network (CNN) [20, 21], have been utilized to improve tracking accuracy and speed. In 2010, Bolme et al. [18] proposed a CF-based method called minimum output sum of squared error (MOSSE), which achieves high-speed tracking with reasonable performance. Thereafter, many researchers have further developed the CF framework to achieve state-of-the-art performance. Li and Zhu [19] realized scale estimation and multiple feature integration on the CF framework. Danelljan et al. [22] eliminated the boundary effect by adding spatial regularization to the learned filter at the cost of speed decrease. Galoogahi et al. [23] provided another efficient solution to solve the boundary effect, thereby maintaining real-time speed. Another popular framework is Siamese-based network, which was first introduced by Bertinetto et al. [20]. Subsequently, deeper and wider networks were introduced to improve target representation. Zhang and Peng [21] found that the padding operation in the deeper network induces position bias, interfering with network capability. To address this problem, they improved the tracking performance significantly. Some of the methods perform better scale estimation by predicting segmentation masks rather than bounding boxes [24, 25]. In summary, there have been many efforts in this field. However, target appearance, as the main cue from visible images, is not reliable for tracking in extreme scenarios including low illumination, out-of-view targets, and heavy occlusion. To this end, more complementary cues are added to handle these challenges. A visible camera is assisted by other sensors, such as laser [26], depth [7], thermal [10], radar [27], and audio [28], to satisfy different requirements.

Since 2005, a series of methods using various multi-modal information have been proposed. Song et al. [26] conducted multiple object tracking by using visible and laser data. Kim and Jeon [27] exploited the traditional Kalman filter for multiple object tracking of radar and visible images. Megherbi et al. [28] proposed a tracking method by combining vision and audio information using belief theory. In this study, tracking of RGB-D and RGB-T data, using a portable and affordable binocular camera, is the focus. As shown in Fig. 1, thermal images capture the target temperature and are not sensitive to lighting conditions, rain, or fog, which can make the tracker work day and night. Depth data provide relative distance between targets and camera and can easily detect target occlusion and capture target boundaries. RGB-D data have been used to detect heavily occluded and out-of-view targets. In recent works, this auxiliary modality has achieved 8.5%–31.3% relative improvement in terms of tracking accuracy on corresponding datasets [10, 29–31]. Thus, the complementary characteristics of multi-modal data can significantly improve tracking accuracy and robustness. Lan et al. [32] applied the sparse learning method to RGB-T tracking, thereby removing the cross-modality discrepancy. Li et al. [11] extended the RGB tracker to the RGB-T domain,

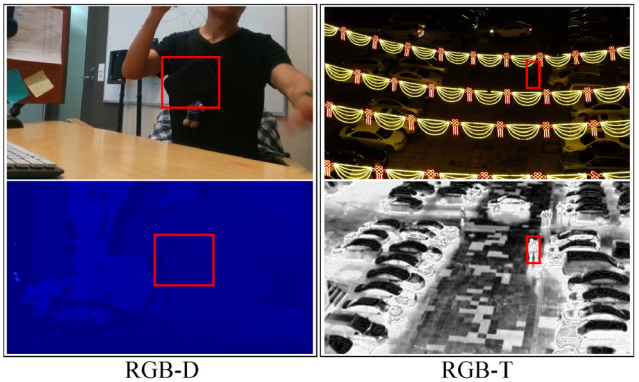

Fig. 1 Visualization of RGB-D and RGB-T pairs.

achieving promising results. Zhang et al. [10] jointly modeled motion and appearance information, to achieve accurate and robust tracking. Kart et al. [7] introduced an effective constraint using a depth map to guide model learning. Liu et al. [33] proposed a mean-shift-based method which transformed the target position to 3D coordinates using RGB and depth images.

2.2 Previous surveys and reviews

In Table 1, existing surveys related to multi-modal processing such as image fusion, object tracking, and multi-modal machine learning are introduced. Some methods focus on specific multi-modal information or single tasks. Cai et al. [36] collected the datasets captured by RGB-D sensors. These datasets are used in many different applications, such as object recognition, scene classification, hand gesture recognition, 3D simultaneous localization and mapping, and pose estimation. Camplani et al. [37] focused on multiple human tracking with RGB-D data and conducted an in-depth review considering different aspects. Ma et al. [39] presented a comprehensive and detailed survey on RGB-T image fusion methods. Recently, a survey on RGB-T object tracking [40] was also presented, analyzing various RGB-T trackers and conducting quantitative analyses on several datasets.

Other surveys provide a general introduction on how to utilize and represent multi-modal information among a series of tasks. Atrey et al. [34] presented a brief introduction on multi-modal fusion methods and analyzed different fusion types in 2010. Walia and Kapoor [35] introduced a general survey on tracking using multiple modality data in 2016. Baltrusaitis et al. [38] provided a detailed review of machine learning methods using multi-modal information.

Various differences and developmental approaches are observed even among the most related works [35, 40]. In this study, first, we conduct a general survey on methods of multi-modal information utilization, especially RGB-D and RGB-T for visual object tracking, under a unified view. In contrast to the survey of Ref. [35], we focus on recent deep-learning-based methods which had not been proposed in 2016. Finally, compared with the literature [40] that only focuses on RGB-T tracking, our study provides a more substantial and comprehensive survey in a larger scope, including RGB-D and RGB-T tracking.

Table 1 Summary of existing surveys in related fields

Index	Year	Reference	Area	Description	Publication
1	2010	[34]	Multi-modal fusion	This paper provides an overview on multi-modal data fusion.	MS
2	2016	[35]	Multi-modal object tracking	This paper provides a general review of both single-modal and multi-modal tracking methods.	AIR
3	2016	[36]	RGB-D dataset	This paper collects popular RGB-D datasets for different applications and provides an analysis on popularity and difficulty.	MTA
4	2017	[37]	RGB-D multiple human tracking	This paper surveys the existing multiple human tracking methods on RGB-D data from two aspects.	IET CV
5	2019	[38]	Multi-modal machine learning	This is a general survey covering representation, translation, and fusion of multi-modal data with regard to various tasks.	TPAMI
6	2019	[39]	RGB-T image fusion	This paper provides a detailed survey on existing methods and applications for RGB-T image fusion.	IF
7	2020	[40]	RGB-T object tracking	A survey of the existing RGB-T tracking methods.	IF

3 Multi-modal visual tracking

This section provides an overview of multi-modal tracking from three aspects: (1) Auxiliary modality: how to utilize the information of auxiliary modality to improve tracking performance. (2) Tracking framework: the types of framework that trackers belong to. (3) Dataset: the utilized datasets in RGB-D and RGB-T tracking tasks. Note that in this study, we mainly focus on visible-thermal (RGB-T) and visible-depth (RGB-D) tracking, considering visible modality as the main modality. Other sources such as thermal and depth, are auxiliary modalities. The taxonomic structure is shown in Fig. 2.

3.1 Auxiliary modality

First, we discuss the purpose of auxiliary modality in multi-modal tracking. There are three main categories: (a) feature learning: the feature representation of the auxiliary modality is extracted to help locate the target; (b) pre-processing: the information from auxiliary modality is used prior to target modeling; and (c) post-processing: the information from auxiliary modality aims to improve the model or refine the bounding box.

3.1.1 Feature learning

Methods based on feature learning extract information from the auxiliary modality through various feature methods, and then adopt modality fusion to combine the data from different sources. Feature learning explicitly utilizes multi-modal information, and most corresponding methods consider the auxiliary modality image as an extra channel of the model. As shown in Fig. 3, these methods can be further categorized into methods based on early fusion (EF) and late fusion (LF) [34, 98]. EF-based methods combine multi-modal information at the feature level using concatenation and summation approaches, whereas LF-based methods model each modality individually and obtain the final result by considering both decisions of modalities.

Early fusion. In EF-based methods, the features extracted from both modalities are first aggregated into a larger feature vector and then sent to the model to locate the target. The workflow of EF-based trackers is shown in the left part of Fig. 3. For most trackers, EF is the primary choice for the multi-modal tracking task, such that visible and auxiliary modalities are treated alike by the same

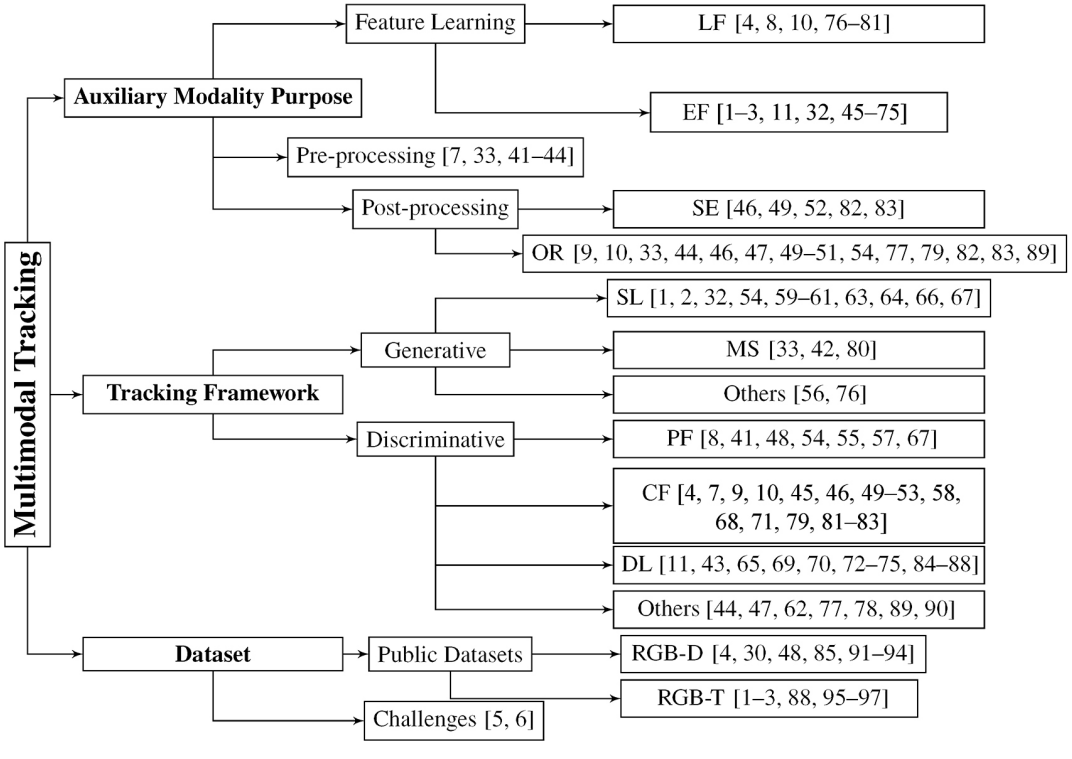

Fig. 2 Structure of three classification methods and algorithms in each category.

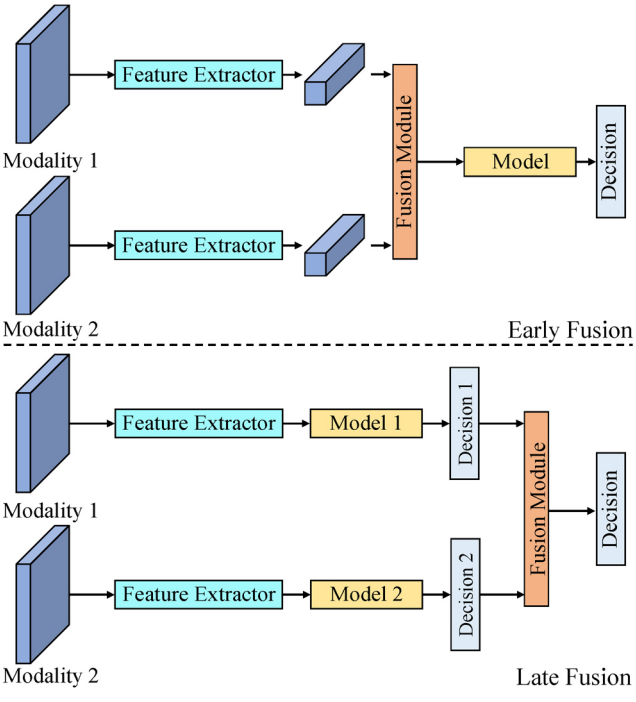

Fig. 3 Workflows of early fusion and late fusion. EF-based methods conduct feature fusion and model them jointly; while LF-based methods aim to model each modality individually and then combine their decisions.

feature extraction methods. Camplani et al. [46] utilized the histogram of oriented gradient (HOG) feature for both visible and depth maps. Kart et al. [50] extracted multiple features to build a robust tracker for RGB-D tracking. Similar methods exist in Refs. [2, 3, 45, 47, 51, 52, 57, 59, 61, 63]. However, the auxiliary modality often indicates information that is different from the visible map. For example, thermal and depth images contain temperature and depth data, respectively. The aforementioned trackers apply feature fusion, ignoring the modality discrepancy, which decreases the tracking accuracy and causes the tracker to easily drift. To this end, some trackers differentiate the heterogeneous modalities by applying different feature methods. In Ref. [48], the gradient feature is extracted in a depth map, while the average color feature is used to represent the target in the visible modality. Meshgi et al. [55] used the raw depth information and many feature methods (HOG, local binary pattern (LBP), and Laplacian of Gaussian (LoG)) for RGB images. In Refs. [32, 60, 67], the HOG and intensity features are used for visible and thermal modalities, respectively. To circumvent the increasing cost involved in feature concatenation and misalignment of multi-modal data, some methods tune the feature representation after feature extraction by pruning [70] or re-weighting [53, 75], which can compress the feature space and exploit cross-modal correlation. In Ref. [70], a feature pruning module is proposed to eliminate noisy and redundant information. Liu et al. [53] introduced a spatial weight to highlight the foreground area. Zhu et al. [75] exploited modality importance using the multi-modal aggregation network.

Late fusion. LF-based methods process both modalities simultaneously, and independent models are built for each modality to make decisions. Subsequently, the decisions are combined by using weighted summation [4, 77, 79, 81], calculating the joint distribution function [8, 76, 80], and conducting multi-step localization [78]. Conaire et al. [76] assumed independence between multi-modal data and multiplied the target likelihoods in both modalities to obtain the result. A similar method has been adopted in the literature [80]. Xiao et al. [4] fused two single-modal trackers via an adaptive weight map. In Ref. [78], data from multiple sources are used step-by-step to locate the target. A rough target position is first estimated by optical flow in the visible domain, and the final result is determined by part-based matching method with RGB-D data.

3.1.2 Pre-processing

The second goal of auxiliary modality is to transform the target into 3D space before target modeling via RGB-D data, using the available depth map. Instead of tracking in the image plane, these types of methods model the target in world coordinates, and 3D trackers are designed [7, 33, 41–44]. Liu et al. [33] extended the classical mean shift tracker to 3D. In Ref. [7], the dynamic spatial constraint generated by the 3D target model enhances the discrimination of DCF trackers in dealing with out-of-view rotation and heavy occlusion. Although significant performance is achieved, the computation cost of 3D reconstruction cannot be neglected. Furthermore, the performance is highly subject to the quality of depth data and the accessibility of mapping functions between 2D and 3D spaces.

3.1.3 Post-processing

Compared with the RGB image that brings more detailed content, the depth image highlights the contour of objects, allowing segmentation of the target from surroundings via depth variance. Inspired

by the nature of the depth map, many RGB-D trackers utilize the depth information to determine the occurrence of an occlusion and estimate the target scale [46, 49, 52, 82].

Occlusion reasoning (OR). Occlusion is a traditional challenge in the tracking task because dramatic appearance variations lead to model drift. Depth cue is a powerful feature for detecting target occlusion; thus, the tracker can apply a global search strategy or model an updating mechanism to avoid learning from the occluded target. In Ref. [46], occlusion is detected when the depth variance is large. Then, the tracker enlarges the search region to detect the re-appearing target. Ding and Song [47] proposed an occlusion recovery method, where a depth histogram is recorded to examine whether an occlusion has occurred. If an occlusion is detected, the tracker locates the occluder and searches a candidate around it. In Ref. [10], Zhang et al. proposed a tracker switcher to detect occlusion based on the template matching method and tracking reliability. The framework of Ref. [10] is shown in Fig. 4. The tracker can dynamically select the information used for tracking between appearance and motion cues, thereby improving the robustness of the tracker significantly.

Scale estimation (SE). SE is an important module in tracking, whereby drift is avoided by obtaining a tight bounding box. CF-based trackers estimate the target scale by sampling the search region in multiple resolutions [99], thereby learning a filter for scale estimation [100] to effectively adapt to the target scale change. Both thermal and depth maps provide clear contour information and a coarse pixel-wise target segmentation map. With such information, the target shape can be effectively estimated [52]. In Ref. [49], the number of scales is adaptively changed to fit the scale variation. SEOH [52] uses space continuity-of-depth information to achieve accurate scale estimation with negligible time cost. The pixels belonging to the target are clustered by applying the K-means method to the depth map, and the sizes of the target and search regions are determined by the clustering result.

3.2 Tracking framework

In this section, multi-modal trackers are categorized based on the methods used in target modeling, including generative and discriminative. The generative framework focuses on directly modeling the representation of the target. During tracking, the target is captured by matching the data distribution in the incoming frame. However, generative methods only learn the representations for the foreground information while ignoring the influence of surroundings which suffer from background cluttering or distractions [101]. In comparison, the discriminative models construct an effective classifier to distinguish the object from the surroundings. The tracker outputs the confidence score of sampled candidates and chooses the best matching patch as the target. Various patch sampling methods are exploited, e.g., sliding window [53], particle filter [41, 48], and Gaussian sampling [11]. A crucial task is to utilize powerful features to represent the target. Emerging convolution networks have enabled more trackers to be built via efficient CNNs. The various frameworks are introduced in the following paragraphs.

3.2.1 Generative methods

Sparse learning (SL). SL has been popular in many tasks including image recognition [102], classification [103], and object tracking [104], among others. In

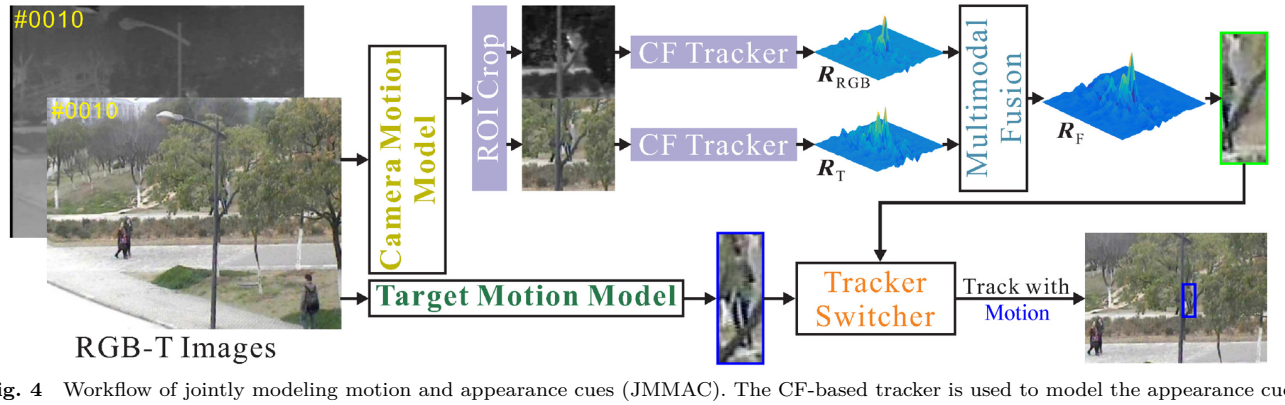

Fig. 4 Workflow of jointly modeling motion and appearance cues (JMMAC). The CF-based tracker is used to model the appearance cue, while both camera and target motion are considered, thereby achieving substantial performance.

SL-based RGB-T trackers, the tracking task can be formulated as a minimization problem for the reconstruction error based on the learned sparse dictionary [1, 32, 59–61, 63, 66, 67]. Lan et al. [32] proposed a unified learning paradigm to learn by modality the target representation, reliability, and classifier, collaboratively. Similar methods are also applied in the RGB-D tracking task. Ma and Xiang [54] constructed an augmented dictionary consisting of target and occlusion templates, thereby achieving accurate tracking even in the presence of heavy occlusion. SL-based trackers achieve promising results at the expense of computation cost. These trackers cannot meet the requirements of real-time tracking.

Mean shift (MS). MS-based methods maximize the similarity between histograms of candidates and the target template, conducting fast local search using the mean shift technique. These methods usually assume that the object overlaps itself in consecutive frames [80]. In Refs. [33, 42], the authors extended the 2D MS method to 3D for RGB-D data. Conaire et al. [80] proposed an MS tracker that uses spatiogram instead of histogram. Compared with discriminative methods, MS-based trackers directly regress the offset of the target, without considering dense sampling. These methods with lightweight features can achieve real-time performance although the performance advantage is not obvious.

Other frameworks. Other generative methods have been applied to tracking tasks. Coraire et al. [76] modeled the tracked object via Gaussian distribution and selected the best-matched patch via a similarity measure. Chen et al. [56] modeled the statistics of each individual modality and the relationship between RGB and thermal data using the expectation maximization algorithm. These methods can model individual or complementary modalities, providing a flexible framework for different scenes.

3.2.2 Discriminative methods

Particle filter. The PF framework is a Bayesian sequential importance sampling technique [105] consisting of two steps, i.e., prediction and updating. In the prediction step, given the state observations $\boldsymbol{z}_{1:t} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_t\}$ during the previous t frames, the posterior distribution of the state $\boldsymbol{x}_t$ is predicted using the Bayesian rule as Eq. (1):

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{z}_{1:t}\right) = \frac{p\left(\boldsymbol{z}_t \mid \boldsymbol{x}_t\right) p\left(\boldsymbol{x}_t \mid \boldsymbol{z}_{1:t-1}\right)}{p\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{1:t-1}\right)} \tag{1}$$

where $p\left(\boldsymbol{x}_t \mid \boldsymbol{z}_{1:t-1}\right)$ is estimated by a set of N particles. Each particle has a weight, w_t^i. In the updating process, w_t^i is updated as

$$w_t^i \propto p\left(\boldsymbol{z}_t \mid \boldsymbol{x}_t = \boldsymbol{x}_t^i\right) \tag{2}$$

In the PF framework, the restrictions of linearity and Gaussianity imposed by the Kalman filter are relaxed, thereby obtaining accurate and robust tracking [8]. Several works have improved the PF method for the multi-modal tracking task. Bibi et al. [41] formulated the PF framework in 3D, considering both representation and motion models and proposed a particle pruning method to boost the tracking speed. Meshgi et al. [55] considered occlusion in the approximation step to improve the occlusion handling of PF. The framework of Ref. [55] is shown in Fig. 5. Liu and Sun [67] proposed a new likelihood function

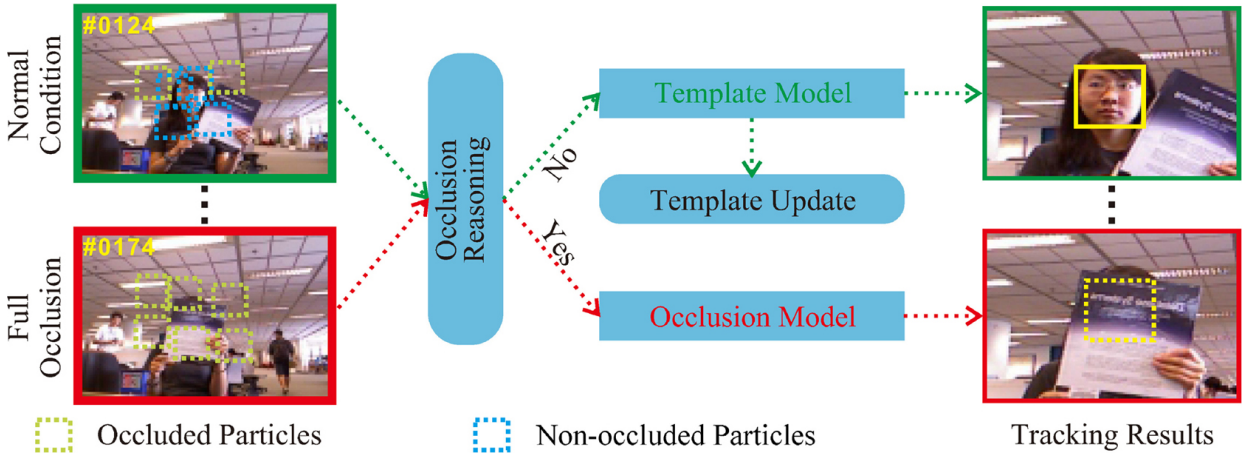

Fig. 5 Occlusion-aware particle filter (OAPF) framework. The particle filter method with occlusion handling is applied, whereby the occlusion model is constructed against the template model. When the target is occluded, the occlusion model is used to predict the position without updating the template model.

for PF to determine the goodness of particles, thereby promoting performance.

Correlation filter. The CF-based tracker learns the discriminative template denoted as CF to represent the target. Then, the online learned filter is used to detect the object in the next frame. As circular convolution can be accelerated in the Fourier domain, these trackers can maintain acceptable accuracy with high speed. In recent years, many CF-based variants have been proposed, such as introducing spatial regularization [106], temporal constraints [107], or discriminative features [108], to increase the tracking performance. The advantages of CF-based trackers have motivated many researchers to build multi-modal trackers within the CF framework. Zhai et al. [68] introduced low-rank constraints to learn a cross-modal correlation filters (CMCF), exploiting the relationship between RGB and thermal data. Hannuna et al. [49] effectively handled the scale change based on the guidance of the depth map. Kart et al. [7] proposed a long-term RGB-D tracker, which is designed based on a channel and spatial reliability discriminative correlation filter (CSRDCF) [109] and applied online 3D target reconstruction to facilitate learning robust filters. The spatial constraint is learned from the 3D model of the target. When the target is occluded, view-specific DCFs are used to robustly localize the target. Camplani et al. [46] improved the CF method for scale estimation and occlusion handling in real time.

Deep learning (DL). CNN is widely used in the tracking task because of its discriminative ability in feature representation. Various networks provide a powerful alternative to the traditional hand-crafted feature, which is the simplest way to utilize CNN. Liu et al. [53] extracted the deep features from VGGNet [110] and used hand-crafted features to learn a robust representation. Li et al. [71] concatenated deep features from visible and thermal images, and then adaptively fused them using the proposed FusionNet to achieve robust feature representation. Some methods aim to learn an end-to-end network for multi-modal tracking. In Refs. [11, 70, 72], a similar framework borrowed from MDNet [111] is applied for tracking different structures to fuse cross-modal data. The MDNet-based framework is shown in Fig. 6. These trackers achieve obvious performance advantages although the speed is low. Zhang et al. [74] proposed an end-to-end real-time RGB-T tracking framework with balanced accuracy. They applied ResNet [112] as the feature extractor and fused RGB and thermal information at the feature level, for target localization and box estimation. Attribute-driven representation network (ADRNet) [29] and challenge-aware tracker (CAT) [113] were proposed to exploit the effectiveness of attribute-specific representation for object modeling. Both trackers contain several individual branches to mine attribute-specific properties, whereby those features are fused to build a powerful representation for accurate tracking.

Based on the vision transformer [114, 115], several

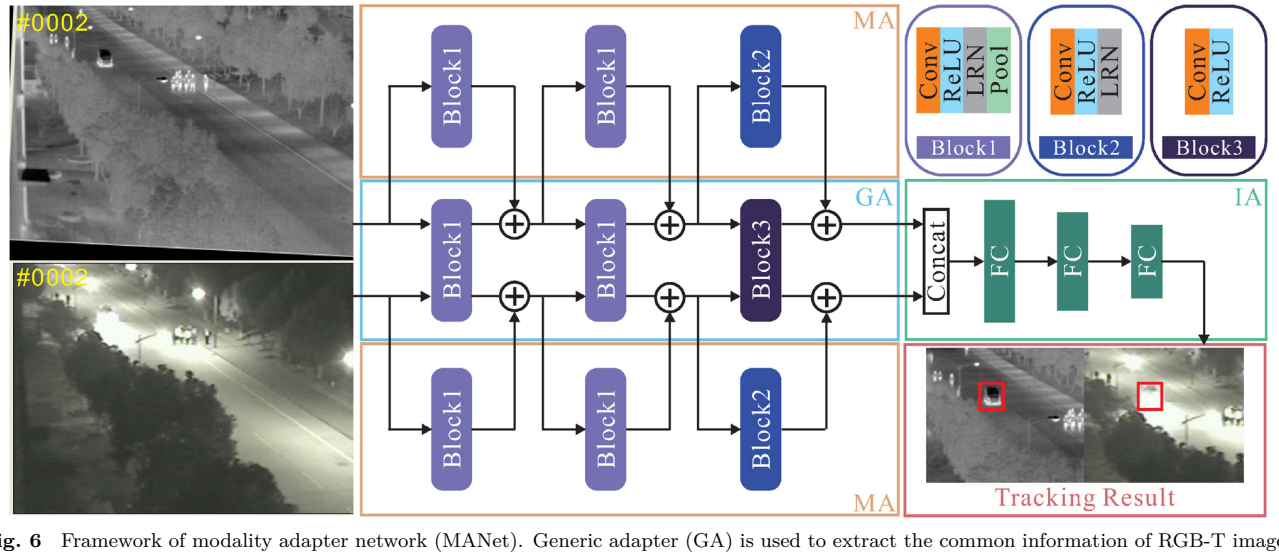

Fig. 6 Framework of modality adapter network (MANet). Generic adapter (GA) is used to extract the common information of RGB-T images. MA exploits the different properties of heterogeneous modalities. Finally, instance adapter (IA) is used to model appearance properties and temporal variations of a certain object.

transformer-based multi-modal trackers have been proposed [85–87] to exploit the complementary cues in both modalities. In Ref. [86], the multi-modal features are first sent to the self-attention module and then cross-attention modules are applied to fuse different branches. Finally, an extra cross-attention module is added to achieve finer fusion. The tracker displays the fusion ability of the transformer framework, achieving state-of-the-art performance. Xiao et al. [87] proposed attribute-based progressive fusion network (APFNet), which aims to fuse attribute-based representation with the proposed enhancement fusion transformer (EFT). The EFT contains three encoders and two decoders, with the encoder enhancing the corresponding input features and the decoder performing the interactive enhancements between aggregated and modality-specific features. As for RGB-D tracking, the SPT tracker was proposed to extract unimodal features, and feature fusion was realized via transformer architecture. The tracker is based on STARK [115], which introduces a transformer encoder for fusing visible and depth information.

Other frameworks. Some methods use an explicit template matching method to localize the object. These methods find the best-matched candidate by capturing the target in frames through a pre-defined matching function [44, 78]. Ding and Song [47] learned a Bayesian classifier and considered the candidate with the maximal score as the target location, thereby reducing model drift. In Ref. [90], a structured support vector machine (SVM) [116] is learned by maximizing a classification score, preventing labeling ambiguity in the training process.

4 Datasets

With the emergence of multi-modal tracking methods, several datasets and challenges for RGB-D and RGB-T tracking have been released. We summarize the available datasets in Table 2.

4.1 Public dataset

4.1.1 RGB-D dataset

In 2012, a small-scale dataset called BoBoT-D [48] was constructed, consisting of five RGB-D video sequences captured by the Kinect V1 sensor. Both overlap and hit rate are used for evaluation, to indicate the mean overlap between result and ground truth and percentage of frames with overlap larger than 0.33. Song and Xiao [93] proposed the well-known Princeton tracking benchmark (PTB) of 100 high-diversity RGB-D videos, of which five are used for validation and others without available ground truth are used for testing. The dataset is for testing tracker performance on occlusion handling and target re-detection. The PTB dataset contains 11 annotations, which are separated into five categories including target type, target size, movement, occlusion, and motion type. Two metrics are used to evaluate the tracking performance: center position error (CPE) and success rate (SR). CPE measures the Euclidean distance between centers of the result and ground truth, whereas SR denotes the average intersection over union (IoU) during all frames and is defined as

$$\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} u_i,\ u_i = \begin{cases} 1, & \mathrm{IoU}(\mathrm{bb}_i, \mathrm{gt}_i) > t_{\mathrm{sr}} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathrm{IoU}(\cdot,\cdot)$ denotes the IoU between the bounding box bb_i and ground truth gt_i in the i-th frame. If

Table 2 Summary of multi-modal tracking datasets

	Name	Seq. Num.	Total frames	Min. frames	Max. frames	Attr.	Resolution	Metrics	Year
RGB-D	PTB	100	21.5k	40	0.90k	11	640 × 480	CPE, SR	2013
	STC	36	18.4k	130	0.7k	10	640 × 480	SR, Acc., Fail.	2018
	CTDB	80	101.9k	400	2.5k	13	640 × 360	F-score, Pr, Re	2019
	DepthTrack	200	294.5k	143	4.0k	15	640 × 360	F-score, Pr, Re	2021
	RGBD1K	1050	2.5M	500	3.0k	15	640 × 360	F-score, Pr, Re	2022
RGB-T	OTCBVS	6	7.2k	600	2.3k	—	320 × 240	—	2007
	LITIV	9	6.3k	300	1.2k	—	320 × 240	—	2012
	GTOT	50	7.8k	40	0.3k	7	384 × 288	SR, PR	2016
	RGBT210	210	104.7k	40	4.1k	12	630 × 460	SR, PR	2017
	RGBT234	234	116.6k	40	8.1k	12	630 × 460	SR, PR, EAO	2019
	VOT-RGBT	60	20.0k	40	1.3k	5	630 × 460	EAO	2019
	LasHeR	1224	734.8k	57	12.8k	19	640 × 480	SR, PR	2021
	VTUAV	500	1.7M	196	27.2k	13	1920 × 1080	SR, PR	2022

IoU is larger than the threshold t_{sr}, the target is considered to be successfully tracked. The final rank of the tracker is determined by Avg. Rank, which is defined as the average ranking of SR in each attribute. The STC dataset [4] consists of 36 RGB-D sequences and covers some extreme tracking circumstances, such as outdoor and night scenes. This dataset is captured by still and moving ASUS Xtion RGB-D cameras to evaluate tracking performance under conditions of arbitrary camera motion. A total of 10 attributes are labeled to thoroughly analyze dataset bias. The illumination of PTB and STC datasets are shown in Fig. 7. The attribute description is provided in the Electronic Supplementary Material (ESM).

The trackers are measured by using both success rate (SR) and visual object tracking (VOT) protocols. The VOT protocol evaluates the tracking performance in terms of accuracy and failure. Accuracy (Acc.) considers the IoU between ground truth and bounding box, and failure (Fail.) keeps track of zero overlap, whereby the tracker is re-initialized when there is no overlap using the ground truth, and tracking is continued. Color-and-depth general visual object tracking benchmark (CTDB) [94], proposed in 2019, contains 80 short-term and long-term videos. The target is frequently out-of-view and occluded, and the tracker needs to handle both tracking and re-detection cases. The metrics used for evaluation are precision (Pr), recall (Re), and overall F-score [117]. Precision and recall are defined as Eqs. (4) and (5)

$$\mathrm{Pr} = \frac{\sum_{i=1}^{N} u_i}{\sum_{i=1}^{N} s_i},\ u_t = \begin{cases} 1, & \mathrm{bb}_i \ \text{exists} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\mathrm{Re} = \frac{\sum_{i=1}^{N} u_i}{\sum_{i=1}^{N} g_i},\ g_t = \begin{cases} 1, & \mathrm{gt}_i \ \text{exists} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where u_i is defined in Eq. (3). The F-score combines both precision and recall through

$$\text{F-score} = \frac{2\mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{6}$$

In 2021, Yan et al. [30] proposed an RGB-D dataset for long-term tracking, namely DepthTrack which contains 65 unique object categories and consists of 100 training and 50 test sequences. DepthTrack is captured with a mid-price RGB-D sensor (Intel RealSense 415), and the images are stored using 640 × 360 resolution with a framerate of 30 fps. All the frames are annotated using 15 attributes extended from CTDB. The evaluation metrics are the same as those used for CTDB. Recently, a large-scale RGB-D tracking dataset (RGBD1K [85]) was released. RGBD1K contains 1050 sequences (1000 sequences for training and 50 sequences for evaluation) and more than 100 object categories, significantly enlarging both the number of sequences and object categories. The training data are sparsely annotated, with bounding boxes used for annotation of the first 600 frames for each sequence. The same attribute annotation and evaluation metrics of DepthTrack are applied.

4.1.2 RGB-T dataset

In previous years, two RGB-T people detection datasets were used for tracking. The OTCBVS dataset [95] has six grayscale-thermal video clips captured from two outdoor scenes. The LITIV dataset [96] contains nine sequences, considering the effect of illumination in indoor captures. These datasets with limited sequences and low diversity have been depreciated. In 2016, for RGB-T tracking, Li et al. constructed the GTOT dataset which consists of 50 grayscale-thermal sequences. The data are mainly captured by a surveillance camera, covering different scenarios and conditions, such as dark night, rainy day, high illumination, etc. A new attribute for RGB-T tracking is the thermal crossover (TC) label, which indicates that the target has temperature similar to that of the background. Inspired by Refs. [118, 119], GTOT adopts success rate (SR) and precision rate (PR) for evaluation. PR denotes the percentage

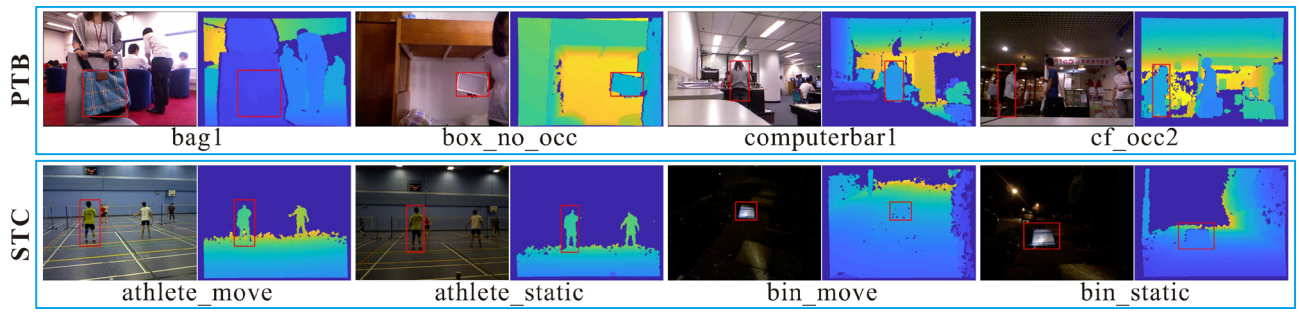

Fig. 7 Examples in RGB-D tracking datasets (PTB and STC datasets).

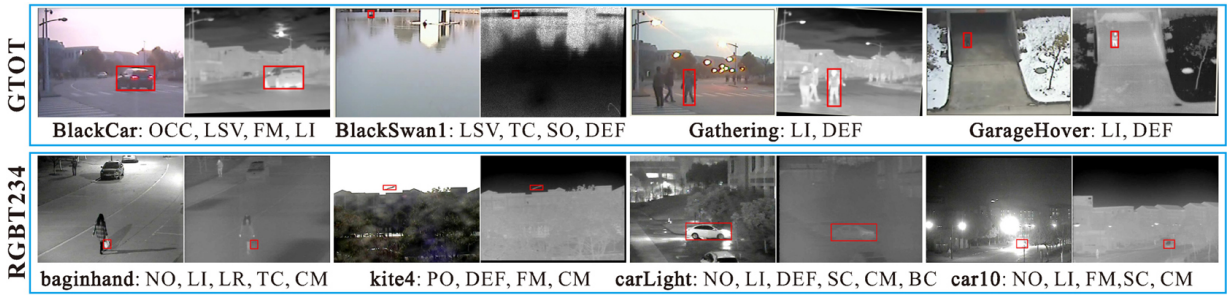

Fig. 8 Examples and corresponding attributes in GTOT and RGBT234 tracking datasets.

of frames with CPE smaller than a threshold $t_{\rm pr}$, which is set to five in GTOT to evaluate small targets. Li et al. [2] proposed a large-scale RGB-T tracking dataset, namely RGBT210, which contains 210 videos and 104.7k image pairs. The dataset includes more challenging cases, such as thermal crossover, small object, and fast motion, allowing for a comprehensive evaluation of the tracker. This dataset also extends the number of attributes to 12. The detailed description of attributes can be found in the ESM. The metrics are the same as those of GTOT, except that $t_{\rm pr}$ is normally set to 20. In 2019, the researchers expanded the RGBT210 dataset and proposed RGBT234 [3], which provides ground truths for each modality. Furthermore, apart from SR and PR, expected average overlap (EAO) is used for evaluation, to combine accuracy and failure in a principled manner. In 2021, Li et al. [97] proposed a large-scale RGB-T tracking dataset, LasHeR, consisting of 1224 visible and thermal infrared video pairs with more than 730k frame pairs in total. Compared with GTOT and RGBT234, LasHeR covers various object categories, viewpoints, scenes, and environmental factors across seasons such as the weather, day, and night. The number of attributes is extended to 19, including several new challenges into the data collection. MSR and MPR are used for evaluation. Recently, Zhang et al. [88] proposed the VTUAV benchmark, which is designed for RGB-T unmanned aerial vehicle (UAV) tracking. VTUAV contains nearly 1.7 million high-resolution RGB-T image pairs with 500 sequences for unveiling the power of RGB-T tracking. The benchmark can be used for evaluating both short-term tracking, long-term tracking, and tracking with segmentation. Furthermore, sequence- and frame-level attribute annotation is provided in VTUAV for 13 typical challenges, to exploit the power of challenge-aware trackers.

4.2 Challenges for multi-modal tracking

Since 2019, both RGB-D and RGB-T challenges have been held by the VOT Committee [5, 6]. In the RGB-D challenge, trackers are evaluated on the CDTB dataset [94] using the same evaluation metrics. All the sequences are annotated on the basis of five attributes, namely, occlusion, dynamics change, motion change, size change, and camera motion. The RGB-T challenge constructs the dataset as a subset of RGBT234 with slight changes in ground truth. The dataset consists of 60 RGB-T public videos and 60 sequestered videos. Compared with RGBT234, VOT-RGBT utilizes different evaluation metrics, e.g., EAO, to measure trackers. In VOT2019-RGBT, trackers need to be re-initialized upon detecting tracking failure (the overlap between bounding box and ground truth is zero). In addition, VOT2020-RGBT introduces a new anchor mechanism to avoid a causal correlation between the first reset and subsequent ones [5] instead of the re-initialization mechanism.

5 Experiments

In this section, we conduct an analysis on both public datasets and challenges for an overall comparison, attribute-based comparison, and speed. For a fair comparison on speed, we consider the device (CPU or GPU), platform (M: MATLAB, MCN: Matconvnet, P: Python, and PT: PyTorch), and settings (detailed information on CPU and GPU). The available code and detailed description of trackers are listed in the ESM.

5.1 Experimental comparison on RGB-D datasets

5.1.1 Overall comparison

PTB provides a website[①] for an online comprehensive

① http://tracking.cs.princeton.edu/

evaluation of RGB and RGB-D methods. We collected the results of 14 RGB-D trackers from the website and sorted them based on rank. The results are shown in Table 3. The Avg. Rank, SR and corresponding rank of each attribute are listed. The Avg. Rank is calculated by averaging the rankings of all attributes. According to Table 3, OTR, which is based on the CF framework without deep features, achieves the best performance among all competitors. The reason for the promising result is that 3D construction provides a useful constraint for filter learning. The same conclusion is obtained for CA3DMS and 3DT, which construct a 3D model to locate the target via mean-shift and sparse learning methods. These trackers with traditional features are competitive with deep trackers. DL-based trackers (WCO, TACF, and CSR-RGBD) achieve substantial performance, indicating the successful discrimination of deep features. CF-based trackers are the most widely-applied framework, and the results differ such that the trackers based on original CF methods (DMKCF, DSKCF, and DSOH) perform significantly worse than those developed on improved CF (OTR, WCO, and TACF). OTOD based on point cloud does not exploit the effectiveness of CNN and attains the 10th rank on the PTB dataset.

5.1.2 Attribute-based comparison

PTB provides 11 attributes and five features for comparison. CF-based trackers, including OTR, WCO, TACF, CSR-RGBD, and CCF, do not perform well in tracking animals. As animal movements are fast and irregular, these online trackers are fragile to drift. When the target is small in size, CF can provide precise tracking results. The occlusion handling mechanism contributes greatly to videos with target occlusion. The 3D mean shift method shows obvious advantages in tracking targets with rigid shape and no occlusion. OAPF achieves an above-average performance in tracking small objects, which indicates the effectiveness of the scale estimation strategy.

5.1.3 Speed analysis

The speed report of RGB-D trackers are listed in Table 4. Most of the trackers cannot meet real-time tracking requirements. Trackers based on the improved CF framework (OTR [7], DMKCF [50], CCF [83], WCO [53], and TACF [51]), are constrained by their own speed. Two real-time trackers (DSKCF [49] and DSOH [46]) have the original CF architecture. The transformer-based tracker SPT, also achieves favorable performance and real-time speed.

5.2 Experimental comparison of RGB-T datasets

We selected 19 trackers as our baseline to perform an overall comparison of the GTOT and RGBT234 datasets. As the code has been released for only part of the trackers (JMMAC, MANet, mfDiMP), we ran these trackers on the two datasets and recorded the performance of the other trackers based on the reports in the original papers. The overall results are shown in Table 5.

5.2.1 Overall comparison

All high-performance trackers are equipped with learned deep features. The transformer-based tracker

Table 3 Experimental results on the PTB dataset. The top three results are in **red**, blue, and green fonts

Algorithm	Target type			Target size		Movement		Occlusion		Motion type	
	Human	Animal	Rigid	Large	Small	Slow	Fast	Yes	No	Passive	Active
OTR	77.3	68.3	81.3	76.5	**77.3**	81.2	**75.3**	71.3	84.7	**85.1**	73.9
WCO	78.0	67.0	80.0	76.0	75.0	78.0	73.0	66.0	86.0	85.0	**82.0**
TACF	76.9	64.7	79.5	77.2	74.0	78.5	74.1	68.3	85.1	83.6	72.3
CA3DMS	66.3	74.3	**82.0**	73.0	74.2	79.6	71.4	63.2	**88.1**	82.8	70.3
CSR-rgbd	76.6	65.2	75.9	75.4	73.0	79.6	71.8	70.1	79.4	79.1	72.1
3DT	**81.4**	64.2	73.3	79.9	71.2	75.1	74.9	**72.5**	78.3	79.0	73.5
DLST	77.0	69.0	73.0	**80.0**	70.0	73.0	74.0	66.0	85.0	72.0	75.0
OAPF	64.2	**84.8**	77.2	72.7	73.4	**85.1**	68.4	64.4	85.1	77.7	71.4
CCF	69.7	64.5	81.4	73.1	72.9	78.4	70.8	65.2	83.7	84.4	68.7
OTOD	72.0	71.0	73.0	74.0	71.0	76.0	70.0	65.0	82.0	77.0	70.0
DMKCF	76.0	58.0	76.7	72.4	72.8	75.2	71.6	69.1	77.5	82.5	68.9
DSKCF	70.9	70.8	73.6	73.9	70.3	76.2	70.1	64.9	81.4	77.4	69.8
DSOH	67.0	61.2	76.4	68.8	69.7	75.4	66.9	63.3	77.6	78.8	65.7
DOHR	45.0	49.0	42.0	48.0	42.0	50.0	43.0	38.0	54.0	54.0	41.0

Table 4 Speed analysis of RGB-D trackers

Tracker	Speed	Device	Platform	Setting
SPT	25.0	GPU	PT	I9@3.6 GHz, RTX3090
OTR	2.0	CPU	M	I7@3.6 GHz
WCO	9.5	GPU	M & MCN	I7@3.4 GHz, GTX Titan
TACF	13.1	GPU	M & MCN	I7@4.0 GHz
CA3DMS	63	CPU	C++	I7@3.6 GHz
DLST	4.8	CPU	M	I5@3.10 GHz
OAPF	0.9	CPU	M	—
CCF	6.3	CPU	M	I7@3.4 GHz
DMKCF	8.3	CPU	M	I7@3.6 GHz
DSKCF	40	CPU	M & C++	I7@3.10 GHz
DSOH	40	CPU	M	I7@3.10 GHz

LRMWT outperformed other trackers with a large margin in terms of SR on both datasets, demonstrating the potential of transformer tracking. HMFT, which is trained on the large-scale benchmark, also showed comparable results. The following trackers which achieve satisfactory results are mainly MDNet variants (CMPP, MaCNet, TODA, DAFNet, MANet, DAPNet, and FANet). Compared with CF trackers, MDNet-based trackers can provide precise target position with higher PR, but are inferior to the CF framework in scale estimation, as reflected by SR. Trackers with sparse learning techniques (CSR, SGT) are better than the L1-PF based on particle filtering. Although mfDiMP utilizes a state-of-the-art backbone, the performance is not positive. The main reason may be that mfDiMP utilizes different training data generated by image translation methods [120], which may cause a gap between existing real RGB-T data.

5.2.2 Attribute-based comparison

We conducted attribute-based comparisons on RGBT234, as shown in Fig. 9. The improved MDNet-based trackers achieve satisfactory performance in the case of low-resolution, deformation, background clutter, fast motion, and thermal crossover. Modeling both camera motion and target motion, JMMAC provides strengths in camera movement and partial occlusion; however, the performance in tracking fast-moving targets is inferior. This condition may result from CF-based trackers having a fixed search region. When the target moves outside the region, the target cannot be detected, thereby causing tracking failure. CMPP, which exploits inter-modal and cross-modal correlations, provides great advantages for low illumination, low resolution, and thermal crossover. The appearance of these attributes are not reliable, and CMPP can eliminate the gap between heterogeneous modalities. The detailed attribute-based comparison can be found in the ESM.

5.2.3 Speed analysis

For tracking speed, the platforms and settings are listed in Table 5 for a fair comparison. Recently-proposed trackers (LRMWT, HMFT, ADRNet) tend to be real-time. DAFNet is based on a real-time MDNet variant, achieving fast tracking with 23.0 fps. Although mfDiMP is equipped with ResNet-101, it was the second fastest tracker because most of the network is trained offline without online tuning.

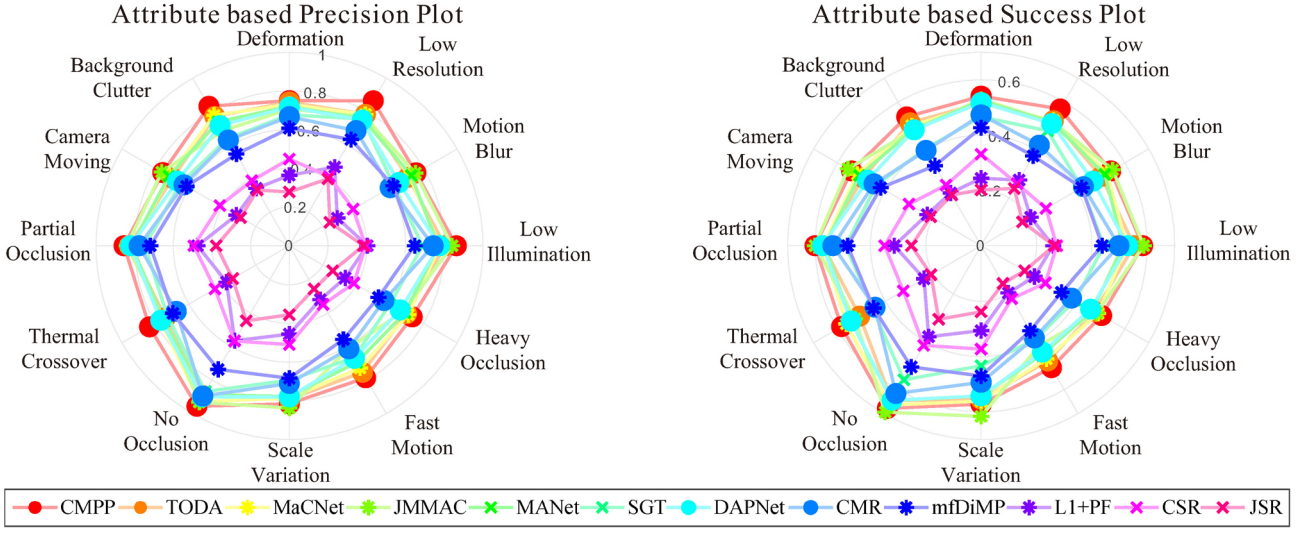

Fig. 9 Attribute-based comparison on RGBT234.

Table 5 Experimental results on the GTOT and RGBT234 datasets

Tracker	GTOT (SR/PR)	RGBT234 (SR/PR)	Speed	Device	Platform	Setting
LRMWT	**75.3**/91.1	**61.6**/82.5	24.6	GPU	PT	GTX 1080Ti
HMFT	74.9/91.2	56.8/78.8	30.2	GPU	PT	RTX Titan
ADRNet	73.9/90.4	57.1/80.9	25.0	GPU	PT	RTX 2080Ti
CMPP	73.8/**92.6**	57.5/82.3	1.3	GPU	PT	RTX 2080Ti
APFNet	73.7/90.5	57.9/**82.7**	—	GPU	PT	GTX 1080Ti
JMMAC	73.2/90.1	57.3/79.0	4.0	GPU	MCN	RTX 2080Ti
CAT	71.7/88.9	56.1/80.4	20.0	GPU	PT	RTX 2080Ti
MaCNet	71.4/88.0	55.4/79.0	0.8	GPU	PT	GTX 1080Ti
TODA	67.7/84.3	54.5/78.7	0.3	GPU	PT	GTX 1080Ti
DAFNet	71.2/89.1	54.4/79.6	23.0	GPU	PT	RTX 2080Ti
MANet	72.4/89.4	53.9/77.7	1.1	GPU	PT	GTX 1080Ti
DAPNet	70.7/88.2	53.7/76.6	—	GPU	PT	GTX 1080Ti
FANet	69.8/88.5	53.2/76.4	1.3	GPU	PT	GTX 1080Ti
CMR	64.3/82.7	48.6/71.1	8.0	CPU	C++	—
SGT	62.8/85.1	47.2/72.0	5.0	CPU	C++	—
mfDiMP	49.0/59.4	42.8/64.6	18.6	GPU	PT	RTX 2080Ti
CSR	—	32.8/46.3	1.6	CPU	M & C++	—
L1-PF	42.7/55.1	28.7/43.1	—	—	—	—
JSR	43.0/55.3	23.4/34.3	0.8	CPU	M	—

Other trackers are constrained by their low speed and cannot be utilized in real-time applications.

5.3 Challenge results on VOT2019-RGBD

The challenge results are listed in Table 6. Both the original RGB tracker without depth information and RGB-D tracker are merged for evaluation. The trackers that secured the top three ranks on F-score, precision, and recall, were designed with the same components in the same framework. Unlike the PTB dataset, DL-based methods perform well on VOT-RGBD19, which is attributed to these trackers utilizing deeper networks and large-scale visual datasets for offline training. For instance, the original RGB tracker with DL framework also achieves excellent performance. Occlusion handling is another necessary component of the high-performance tracker because VOT2019-RGBD focuses on the long-term tracking of frequently reappearing and out-of-view targets. Thus, most of these trackers are equipped with a re-detection mechanism. The CF framework (FuCoLoT, OTR, CSR-RGBD, and ECO) does not perform well, which may stem from online updating using occlusion patches that degrade model discrimination.

Table 6 Challenge results on VOT2019-RGBD dataset

Tracker	Modality	F-score	Precison	Recall
SiamDW-D	RGB-D	**0.681**	**0.677**	0.685
ATCAIS	RGB-D	0.676	0.643	**0.712**
LTDSE-D	RGB-D	0.658	0.674	0.643
SiamM-D	RGB-D	0.455	0.463	0.447
MDNet	RGB	0.455	0.463	0.447
MBMD	RGB	0.441	0.454	0.429
FuCoLoT	RGB	0.391	0.459	0.340
OTR	RGB-D	0.336	0.364	0.312
SiamFC	RGB	0.333	0.356	0.312
CSR-rgbd	RGB-D	0.332	0.375	0.397
ECO	RGB	0.329	0.317	0.342
CA3DMS	RGB	0.271	0.284	0.259

5.4 Challenge results on VOT2019-RGBT

For the VOT2019-RGBT dataset shown in Table 7, JMMAC which exploits both appearance and motion cues exhibits high accuracy and robust performance, obtaining the highest EAO with a large margin. Early fusion is the primary method in RGB-T fusion, whereas the late fusion method JMMAC, which is not fully utilized, has great potential in improving tracking accuracy and robustness. All top six trackers

Table 7 Challenge results on the VOT2019-RGBT dataset

Tracker	Modality	EAO	Acc.	R.
JMMAC	RGB-T	**0.4826**	**0.6649**	**0.8211**
SiamDW-T	RGB-T	0.3925	0.6158	0.7839
mfDiMP	RGB-T	0.3879	0.6019	0.8036
FSRPN	RGB-T	0.3553	0.6362	0.7069
MANet	RGB-T	0.3463	0.5823	0.7010
MPAT	RGB	0.3180	0.5723	0.7242
CISRDCF	RGB-T	0.2923	0.5215	0.6904
GESBTT	RGB-T	0.2896	0.6163	0.6350

are equipped with CNN as the feature extractor, indicating the power of CNN. SiamDW, which uses a Siamese network, is a general method that performs well in both RGB-D and RGB-T tasks. ATOM variants (mfDiMP and MPAT) are used to handle RGB-T tracking.

6 Further prospects

6.1 Model design

6.1.1 Multi-modal fusion

Compared with tracking unimodal data, multi-modal tracking can easily exploit the powerful data fusion mechanism. Existing methods mainly focus on feature fusion, as the effectiveness of other fusion types has not been explored. Compared with early fusion, late fusion eliminates the bias that exists in learning heterogeneous features from different modalities. Another advantage of late fusion is that various methods can be utilized to model each modality independently. As a better choice for multi-modal tracking, the hybrid fusion method which combines early and late fusion strategies, has been used in image segmentation [121] and sports video analysis [122].

6.1.2 Specific network for auxiliary modality

As a gap exists between different modalities, and semantic information is heterogeneous, traditional methods use different features to extract more useful data [48, 60, 67]. Although sufficient studies have been conducted on network architectures for visible image analysis, the specific structure of depth and thermal maps has not been deeply explored. Thus, DL-based methods [11, 69, 70, 74] trade data in auxiliary modality as an additional dimension of the RGB image with the same network architecture (e.g., VGGNet and ResNet) and extract the feature at the same level (layer). A crucial task is to design a network for multi-modal data processing. Since 2017, the AutoML method, especially neural architecture search (NAS), has been popular in automatically designing the architecture, obtaining highly competitive results in many areas such as image classification [123] and recognition [124]. However, for multi-modal tracking, the researchers do not pay as much attention to the NAS method, which is a good direction to explore.

6.1.3 Multi-modal tracking with real-time speed

The additional modality increases computation, causing difficulty for existing tracking frameworks to achieve the requirements of real-time performance. Thus, a speed-up mechanism such as feature selection [70] or knowledge distillation technology needs to be designed. Furthermore, Huang et al. [125] proposed a trade-off method, whereby the agent decides on the more suitable layer for accurate localization, which provides a speed boost of 100 times.

6.2 Dataset construction

6.2.1 Large-scale dataset with high diversity

With the emergence of deep neural networks, more powerful methods are equipped with CNN and transformer to achieve accurate and robust performance, thereby requiring numerous training samples to unveil the power of large models. Recent benchmarks [1, 3, 88] mainly focus on a single application, such as surveillance or drone tracking, and the target category is also limited. With the popularity of multi-modal cameras, a new large-scale dataset with high diversity should be set up to promote the development of multi-modal tracking.

6.2.2 Modality registration

As multi-modal data are captured by different sensors and the binocular camera has a parallax error that cannot be ignored when the target is small and resolution is low, registering the data in spatial and temporal dimensions is essential. In the VOT-RGBT challenge, the dataset ensures the precise annotation in infrared modality, and the tracker handles the misalignment of the RGB image. We state that the image pre-registration process which involves cropping the shared visual field and applying an image registration method, is necessary during dataset construction.

6.2.3 Metrics for robustness evaluation

In some extreme scenes and weather conditions, such as rain, low illumination, and hot sunny days, visible or thermal sensors cannot provide meaningful data. The depth camera cannot precisely estimate distance when the object is far from the sensor. Therefore, a robust tracker needs to avoid tracking failure when any of the modality data are unavailable for a certain period. To handle this case, both complementary and discriminative features have to be applied in

localization. However, none of the datasets measure tracking robustness in the presence of missing data or adversarial attacks [126, 127]. Thus, a new evaluation metric must be considered to track robustness.

6.2.4 Tracking in specific scenes and applications

Recent datasets are designed for generic object tracking of various target categories and scenes. However, these datasets cannot fully exploit the potential of an auxiliary modality for existing trackers with unimodal input to achieve reasonable results. Tracking in specific scenes (dark night, rain, and crowds) requires trackers to have stronger capabilities for information fusion and modality switching to facilitate the development of multi-modal tracking frameworks.

7 Conclusions

In this study, an in-depth review of multi-modal tracking is provided. First, we summarize multi-modal trackers into a unified framework, and analyze them from different perspectives, including auxiliary modality, purpose, and tracking framework. Then, a detailed introduction is presented on the datasets for multi-modal tracking along with the corresponding metrics. Furthermore, a comprehensive comparison on five popular datasets is conducted, and the effectiveness of trackers belonging to various types are analyzed from the perspectives of overall performance, attribute-based performance, and speed. Finally, as an emerging field, several possible directions are identified to facilitate the improvement of multi-modal tracking. The comparison results and analysis will be available at https://github.com/zhang-pengyu/Multimodal_tracking_survey.

Acknowledgements

The study was supported in part by National Natural Science Foundation of China (Nos. U23A20384 and 62022021), in part by Joint Fund of Ministry of Education for Equipment Pre-research (No. 8091B032155), in part by the National Defense Basic Scientific Research Program (No. WDZC20215250205), and in part by Central Guidance on Local Science and Technology Development Fund of Liaoning Province (No. 2022JH6/100100026).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

Electronic supplementary material is available in the online version of this article at `https://doi.org/10.1007/s41095-023-0345-5`.

References

[1] Li, C. L.; Cheng, H.; Hu, S. Y.; Liu, X. B.; Tang, J.; Lin, L. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing* Vol. 25, No. 12, 5743–5756, 2016.

[2] Li, C. L.; Zhao, N.; Lu, Y. J.; Zhu, C. L.; Tang, J. Weighted sparse representation regularized graph learning for RGB-T object tracking. In: Proceedings of the 25th ACM International Conference on Multimedia, 1856–1864, 2017.

[3] Li, C. L.; Liang, X. Y.; Lu, Y. J.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition* Vol. 96, 106977, 2019.

[4] Xiao, J. J.; Stolkin, R.; Gao, Y. Q.; Leonardis, A. Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE Transactions on Cybernetics* Vol. 48, No. 8, 2485–2499, 2018.

[5] Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; Kämäräinen, J. K.; Danelljan, M.;, Lukežič, A.; Drbohlav, O.; He, L.; et al. The eighth visual object tracking VOT2020 challenge results. In: *Computer Vision – ECCV 2020 Workshops Lecture Notes in Computer Science, Vol. 12539*. Bartoli, A.; Fusiello, A. Eds. Springer Cham, 547–601, 2020.

[6] Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L.; Fernandez, G.; Vojir, T.; Hager, G.; Nebehay, G.; Pflugfelder, R. The visual object tracking VOT2015 challenge results. In: Proceedings of the IEEE International Conference on Computer Vision Workshop, 564–586, 2015.

[7] Kart, U.; Lukežič, A.; Kristan, M.; Kämäräinen, J.-K.; Matas, J. Object tracking by reconstruction with view-specific discriminative correlation filters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1339–1348, 2019.

[8] Cvejic, N.; Nikolov, S. G.; Knowles, H. D.; Loza, A.; Achim, A.; Bull, D. R.; Canagarajah, C. N. The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, 1–7, 2007.

[9] Kart, U.; Kämäräinen, J. K.; Matas, J. How to make an RGBD tracker? In: *Computer Vision – ECCV 2018 Workshops. Lecture Notes in Computer Science, Vol. 11129*. Leal-Taixé, L.; Roth, S. Eds. Springer Cham, 148–161, 2019.

[10] Zhang, P. Y.; Zhao, J.; Bo, C. J.; Wang, D.; Lu, H. C.; Yang, X. Y. Jointly modeling motion and appearance cues for robust RGB-T tracking. *IEEE Transactions on Image Processing* Vol. 30, 3335–3347, 2021.

[11] Li, C. L.; Lu, A. D.; Zheng, A. H.; Tu, Z. Z.; Tang, J. Multi-adapter RGBT tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 2262–2270, 2019.

[12] Weng, S. K.; Kuo, C. M.; Tu, S. K. Video object tracking using adaptive Kalman filter. *Journal of Visual Communication and Image Representation* Vol. 17, No. 6, 1190–1208, 2006.

[13] Kulikov, G. Y.; Kulikova, M. V. The accurate continuous-discrete extended Kalman filter for radar tracking. *IEEE Transactions on Signal Processing* Vol. 64, No. 4, 948–958, 2016.

[14] Yang, C. J.; Duraiswami, R.; Davis, L. Fast multiple object tracking via a hierarchical particle filter. In: Proceedings of the 10th IEEE International Conference on Computer Vision, Vol. 1, 212–219, 2005.

[15] Okuma, K.; Taleghani, A.; de Freitas, N.; Little, J. J.; Lowe, D. G. A boosted particle filter: Multitarget detection and tracking. In: *Computer Vision - ECCV 2004. Lecture Notes in Computer Science, Vol. 3021*. Pajdla, T.; Matas, J. Eds. Springer Berlin Heidelberg, 28–39, 2004.

[16] Zhang, T. Z.; Ghanem, B.; Liu, S.; Ahuja, N. Low-rank sparse learning for robust visual tracking. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7577*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 470–484, 2012.

[17] Zhang, T. Z.; Ghanem, B.; Liu, S.; Ahuja, N. Robust visual tracking via multi-task sparse learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2042–2049, 2012.

[18] Bolme, D.; Beveridge, J. R.; Draper, B. A.; Lui, Y. M. Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2544–2550, 2010.

[19] Li, Y.; Zhu, J. K. A scale adaptive kernel correlation filter tracker with feature integration. In: *Computer Vision - ECCV 2014 Workshops. Lecture Notes in Computer Science, Vol. 8926*. Agapito, L.; Bronstein, M.; Rother, C. Eds. Springer Cham, 254–265, 2015.

[20] Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; Torr, P. H. S. Fully-convolutional Siamese networks for object tracking. In: *Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science, Vol. 9914*. Hua, G.; Jégou, H. Eds. Springer Cham, 850–865, 2016.

[21] Zhang, Z. P.; Peng, H. W. Deeper and wider Siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4586–4595, 2019.

[22] Danelljan, M.; Hager, G.; Khan, F. S.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, 4310–4318, 2015.

[23] Galoogahi, H. K.; Fagg, A.; Lucey, S. Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, 1144–1152, 2017.

[24] Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W. M.; Torr, P. H. S. Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1328–1338, 2019.

[25] Lukezic, A.; Matas, J.; Kristan, M. D3S—A discriminative single shot segmentation tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7131–7140, 2020.

[26] Song, X.; Zhao, H. J.; Cui, J. S.; Shao, X. W.; Shibasaki, R.; Zha, H. B. An online system for multiple interacting targets tracking. *ACM Transactions on Intelligent Systems and Technology* Vol. 4, No. 1, Article No. 18, 2013.

[27] Kim, D. Y.; Jeon, M. Data fusion of radar and image measurements for multi-object tracking via Kalman filtering. *Information Sciences* Vol. 278, 641–652, 2014.

[28] Megherbi, N.; Ambellouis, S.; Colot, O.; Cabestaing, F. Joint audio-video people tracking using belief theory. In: Proceedings of the IEEE Conference on Advanced Video and Signal based Surveillance, 135–140, 2005.

[29] Zhang, P. Y.; Wang, D.; Lu, H. C.; Yang, X. Y. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *International Journal of Computer Vision* Vol. 129, No. 9, 2714–2729, 2021.

[30] Yan, S.; Yang, J. Y.; Kapyla, J.; Zheng, F.; Leonardis, A.; Kamarainen, J. K. DepthTrack: Unveiling the power of RGBD tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10705–10713, 2021.

[31] Gao, S.; Yang, J. Y.; Li, Z.; Zheng, F.; Leonardis, A.; Song, J. K. Learning dual-fused modality-aware

representations for RGBD tracking. In: *Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, Vol. 13808*. Karlinsky, L.; Michaeli, T.; Nishino, K. Eds. Springer Cham, 478–494, 2023.

[32] Lan, X. Y.; Ye, M.; Zhang, S. P.; Yuen, P. Robust collaborative discriminative learning for RGB-infrared tracking. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 32, No. 1, 7008–7015, 2018.

[33] Liu, Y.; Jing, X. Y.; Nie, J. H.; Gao, H.; Liu, J.; Jiang, G. P. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos. *IEEE Transactions on Multimedia* Vol. 21, No. 3, 664–677, 2019.

[34] Atrey, P. K.; Hossain, M. A.; El Saddik, A.; Kankanhalli, M. S. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* Vol. 16, No. 6, 345–379, 2010.

[35] Walia, G. S.; Kapoor, R. Recent advances on multicue object tracking: A survey. *Artificial Intelligence Review* Vol. 46, No. 1, 1–39, 2016.

[36] Cai, Z. Y.; Han, J. G.; Liu, L.; Shao, L. RGB-D datasets using microsoft kinect or similar sensors: A survey. *Multimedia Tools and Applications* Vol. 76, No. 3, 4313–4355, 2017.

[37] Camplani, M.; Paiement, A.; Mirmehdi, M.; Damen, D. M.; Hannuna, S.; Burghardt, T.; Tao, L. L. Multiple human tracking in RGB-depth data: A survey. *IET Computer Vision* Vol. 11, No. 4, 265–285, 2017.

[38] Baltrusaitis, T.; Ahuja, C.; Morency, L. P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41, No. 2, 423–443, 2019.

[39] Ma, J. Y.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Information Fusion* Vol. 45, 153–178, 2019.

[40] Zhang, X. C.; Ye, P.; Leung, H.; Gong, K.; Xiao, G. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion* Vol. 63, 166–187, 2020.

[41] Bibi, A.; Zhang, T. Z.; Ghanem, B. 3D part-based sparse tracker with automatic synchronization and registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1439–1448, 2016.

[42] Gutev, A.; Debono, C. J. Exploiting depth information to increase object tracking robustness. In: Proceedings of the IEEE EUROCON 18th International Conference on Smart Technologies, 1–5, 2019.

[43] Xie, Y. J.; Lu, Y.; Gu, S. RGB-D object tracking with occlusion detection. In: Proceedings of the 15th International Conference on Computational Intelligence and Security, 11–15, 2019.

[44] Zhong, B. N.; Shen, Y. J.; Chen, Y.; Xie, W. B.; Cui, Z.; Zhang, H. B.; Chen, D. S.; Wang, T.; Liu, X.; Peng, S. J.; et al. Online learning 3D context for robust visual tracking. *Neurocomputing* Vol. 151, 710–718, 2015.

[45] An, N.; Zhao, X. G.; Hou, Z. G. Online RGB-D tracking via detection-learning-segmentation. In: Proceedings of the 23rd International Conference on Pattern Recognition, 1231–1236, 2016.

[46] Camplani, M.; Hannuna, S.; Mirmehdi, M.; Damen, D. M.; Paiement, A.; Tao, L. L.; Burghardt, T. Real-time RGB-D tracking with depth scaling kernelised correlation filters and occlusion handling. In: Proceedings of the British Machine Vision Conference, 145.1–145.11, 2015.

[47] Ding, P.; Song, Y. Robust object tracking using color and depth images with a depth based occlusion handling and recovery. In: Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery, 930–935, 2015.

[48] García, G. M.; Klein, D. A.; Stückler, J.; Frintrop, S.; Cremers, A. B. Adaptive multi-cue 3D tracking of arbitrary objects. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 7476*. Pinz, A.; Pock, T.; Bischof, H.; Leberl, F. Eds. Springer Berlin Heidelberg, 357–366, 2012.

[49] Hannuna, S.; Camplani, M.; Hall, J.; Mirmehdi, M.; Damen, D. M.; Burghardt, T.; Paiement, A.; Tao, L. L. DS-KCF: A real-time tracker for RGB-D data. *Journal of Real-Time Image Processing* Vol. 16, No. 5, 1439–1458, 2019.

[50] Kart, U.; Kamarainen, J. K.; Matas, J.; Fan, L. X.; Cricri, F. Depth masked discriminative correlation filter. In: Proceedings of the 24th International Conference on Pattern Recognition, 2112–2117, 2018.

[51] Kuai, Y. L.; Wen, G. J.; Li, D. D.; Xiao, J. J. Target-aware correlation filter tracking in RGBD videos. *IEEE Sensors Journal* Vol. 19, No. 20, 9522–9531, 2019.

[52] Leng, J. X.; Liu, Y. Real-time RGB-D visual tracking with scale estimation and occlusion handling. *IEEE Access* Vol. 6, 24256–24263, 2018.

[53] Liu, W. C.; Tang, X. A.; Zhao, C. L. Robust RGBD tracking via weighted convolution operators. *IEEE Sensors Journal* Vol. 20, No. 8, 4496–4503, 2020.

[54] Ma, Z. A.; Xiang, Z. Y. Robust object tracking with RGBD-based sparse learning. *Frontiers of Information Technology & Electronic Engineering* Vol. 18, No. 7, 989–1001, 2017.

[55] Meshgi, K.; Maeda, S. I.; Oba, S.; Skibbe, H.; Li, Y. Z.; Ishii, S. An occlusion-aware particle filter tracker to

handle complex and persistent occlusions. *Computer Vision and Image Understanding* Vol. 150, 81–94, 2016.

[56] Chen, S. Y.; Zhu, W. J.; Leung, H. Thermo-visual video fusion using probabilistic graphical model for human tracking. In: Proceedings of the IEEE International Symposium on Circuits and Systems, 1926–1929, 2008.

[57] Wu, Y.; Blasch, E.; Chen, G. S.; Bai, L.; Ling, H. B. Multiple source data fusion via sparse representation for robust visual tracking. In: Proceedings of the 14th International Conference on Information Fusion, 1–8, 2011.

[58] Wang, Y. L.; Li, C. L.; Tang, J.; Sun, D. D. Learning collaborative sparse correlation filter for real-time multispectral object tracking. In: *Advances in Brain Inspired Cognitive Systems. Lecture Notes in Computer Science, Vol. 10989*. Ren, J., et al. Eds. Springer Cham, 462–472, 2018.

[59] Lan, X. Y.; Ye, M.; Shao, R.; Zhong, B. N.; Jain, D. K.; Zhou, H. Y. Online non-negative multi-modality feature template learning for RGB-assisted infrared tracking. *IEEE Access* Vol. 7, 67761–67771, 2019.

[60] Lan, X. Y.; Ye, M.; Zhang, S. P.; Zhou, H. Y.; Yuen, P. C. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognition Letters* Vol. 130, 12–20, 2020.

[61] Lan, X. Y.; Ye, M.; Shao, R.; Zhong, B. N.; Yuen, P. C.; Zhou, H. Y. Learning modality-consistency feature templates: A robust RGB-infrared tracking system. *IEEE Transactions on Industrial Electronics* Vol. 66, No. 12, 9887–9897, 2019.

[62] Li, C. L.; Zhu, C. L.; Huang, Y.; Tang, J.; Wang, L. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11217*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 831–847, 2018.

[63] Li, C. L.; Hu, S. Y.; Gao, S. H.; Tang, J. Real-time grayscale-thermal tracking via Laplacian sparse representation. In: *MultiMedia Modeling. Lecture Notes in Computer Science, Vol. 9517*. Tian, Q.; Sebe, N.; Qi, G. J.; Huet, B.; Hong, R.; Liu, X. Eds. Springer Cham, 54–65, 2016.

[64] Li, C. L.; Zhu, C. L.; Zhang, J.; Luo, B.; Wu, X. H.; Tang, J. Learning local-global multi-graph descriptors for RGB-T object tracking. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 29, No. 10, 2913–2926, 2019.

[65] Zhang, H.; Zhang, L.; Zhuo, L.; Zhang, J. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors* Vol. 20, No. 2, 393, 2020.

[66] Li, C. L.; Sun, X.; Wang, X.; Zhang, L.; Tang, J. Grayscale-thermal object tracking via multitask Laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* Vol. 47, No. 4, 673–681, 2017.

[67] Liu, H. P.; Sun, F. C. Fusion tracking in color and infrared images using joint sparse representation. *Science China Information Sciences* Vol. 55, No. 3, 590–599, 2012.

[68] Zhai, S. L.; Shao, P. P.; Liang, X. Y.; Wang, X. Fast RGB-T tracking via cross-modal correlation filters. *Neurocomputing* Vol. 334, 172–181, 2019.

[69] Gao, Y.; Li, C. L.; Zhu, Y. B.; Tang, J.; He, T.; Wang, F. T. Deep adaptive fusion network for high performance RGBT tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 91–99, 2019.

[70] Zhu, Y. B.; Li, C. L.; Luo, B.; Tang, J.; Wang, X. Dense feature aggregation and pruning for RGBT tracking. In: Proceedings of the 27th ACM International Conference on Multimedia, 465–472, 2019.

[71] Li, C. L.; Wu, X. H.; Zhao, N.; Cao, X. C.; Tang, J. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing* Vol. 281, 78–85, 2018.

[72] Yang, R.; Zhu, Y. B.; Wang, X.; Li, C. L.; Tang, J. Learning target-oriented dual attention for robust RGB-T tracking. In: Proceedings of the IEEE International Conference on Image Processing, 3975–3979, 2019.

[73] Zhang, X. M.; Zhang, X. H.; Du, X. D.; Zhou, X. M.; Yin, J. Learning multi-domain convolutional network for RGB-T visual tracking. In: Proceedings of the 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, 1–6, 2018.

[74] Zhang, L. C.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; Shahbaz Khan, F. Multi-modal fusion for end-to-end RGB-T tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 2252–2261, 2019.

[75] Zhu, Y. B.; Li, C. L.; Luo, B.; Tang, J. FANet: Quality-aware feature aggregation network for robust RGB-T tracking. *arXiv preprint* arXiv:1811.09855, 2018.

[76] Conaire, C. O.; O'Connor, N. E.; Cooke, E.; Smeaton, A. F. Comparison of fusion methods for thermo-visual surveillance tracking. In: Proceedings of the 9th International Conference on Information Fusion, 1–7, 2006.

[77] Shi, H. Z.; Gao, C. X.; Sang, N. Using consistency of depth gradient to improve visual tracking in RGB-D sequences. In: Proceedings of the Chinese Automation Congress, 518–522, 2015.

[78] Wang, Q.; Fang, J. W.; Yuan, Y. Multi-cue based tracking. *Neurocomputing* Vol. 131, 227–236, 2014.

[79] Zhang, H.; Cai, M.; Li, J. X. A real-time RGB-D tracker based on KCF. In: Proceedings of the Chinese Control and Decision Conference, 4856–4861, 2018.

[80] Conaire, C. Ó.; O'Connor, N. E.; Smeaton, A. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. *Machine Vision and Applications* Vol. 19, Nos. 5–6, 483–494, 2008.

[81] Luo, C. W.; Sun, B.; Yang, K.; Lu, T. R.; Yeh, W. C. Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. *Infrared Physics & Technology* Vol. 99, 265–276, 2019.

[82] Zhai, Y. Y.; Song, P.; Mou, Z. L.; Chen, X. X.; Liu, X. J. Occlusion-aware correlation particle filter target tracking based on RGBD data. *IEEE Access* Vol. 6, 50752–50764, 2018.

[83] Li, G. Q.; Huang, L.; Zhang, P. C.; Li, Q.; Huo, Y. K. Depth information aided constrained correlation filter for visual tracking. *IOP Conference Series*: *Earth and Environmental Science* Vol. 234, 012005, 2019.

[84] Wang, C. Q.; Xu, C. Y.; Cui, Z.; Zhou, L.; Zhang, T.; Zhang, X. Y.; Yang, J. Cross-modal pattern-propagation for RGB-T tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7062–7071, 2020.

[85] Zhu, X. F.; Xu, T. Y.; Tang, Z. Y.; Wu, Z. C.; Liu, H. D.; Yang, X.; Wu, X. J.; Kittler, J. RGBD1K: A large-scale dataset and benchmark for RGB-D object tracking. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 37, No. 3, 3870–3878, 2023.

[86] Feng, M. Z.; Su, J. B. Learning reliable modal weight with transformer for robust RGBT tracking. *Knowledge-Based Systems* Vol. 249, 108945, 2022.

[87] Xiao, Y.; Yang, M. M.; Li, C. L.; Liu, L.; Tang, J. Attribute-based progressive fusion network for RGBT tracking. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 36, No. 3, 2831–2838, 2022.

[88] Zhang, P. Y.; Zhao, J.; Wang, D.; Lu, H. C.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8876–8885, 2022.

[89] Chen, Y.; Shen, Y. J.; Liu, X.; Zhong, B. N. 3D object tracking via image sets and depth-based occlusion detection. *Signal Processing* Vol. 112, 146–153, 2015.

[90] Li, C. L.; Zhu, C. L.; Zheng, S. F.; Luo, B.; Tang, J. Two-stage modality-graphs regularized manifold ranking for RGB-T tracking. *Signal Processing*: *Image Communication* Vol. 68, 207–217, 2018.

[91] Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research* Vol. 32, No. 11, 1231–1237, 2013.

[92] Liu, J.; Liu, Y.; Cui, Y.; Chen, Y. Q. Real-time human detection and tracking in complex environments using single RGBD camera. In: Proceedings of the IEEE International Conference on Image Processing, 3088–3092, 2013.

[93] Song, S. R.; Xiao, J. X. Tracking revisited using RGBD camera: Unified benchmark and baselines. In: Proceedings of the IEEE International Conference on Computer Vision, 233–240, 2013.

[94] Lukezic, A.; Kart, U.; Kapyla, J.; Durmush, A.; Kamarainen, J. K.; Matas, J.; Kristan, M. CDTB: A color and depth visual object tracking dataset and benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10021, 2019.

[95] Davis, J. W.; Sharma, V. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding* Vol. 106, Nos. 2–3, 162–182, 2007.

[96] Torabi, A.; Massé, G.; Bilodeau, G. A. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding* Vol. 116, No. 2, 210–221, 2012.

[97] Li, C. L.; Xue, W. L.; Jia, Y. Q.; Qu, Z. C.; Luo, B.; Tang, J.; Sun, D. D. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing* Vol. 31, 392–404, 2022.

[98] Ramachandram, D.; Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* Vol. 34, No. 6, 96–108, 2017.

[99] Li, Y.; Zhu, J. K. A scale adaptive kernel correlation filter tracker with feature integration. In: *Computer Vision - ECCV 2014 Workshops. Lecture Notes in Computer Science, Vol. 8926*. Agapito, L.; Bronstein, M.; Rother, C. Eds. Springer Cham, 254–265, 2015.

[100] Danelljan, M.; Hager, G.; Khan, F. S.; Felsberg, M. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 8, 1561–1575, 2017.

[101] Li, X.; Hu, W. M.; Shen, C. H.; Zhang, Z. F.; Dick, A.; Van Den Hengel, A. A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology* Vol. 4, No. 4, Article No. 58, 2013.

[102] Shojaeilangari, S.; Yau, W. Y.; Nandakumar, K.; Li, J.; Teoh, E. K. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Transactions on Image Processing* Vol. 24, No. 7, 2140–2152, 2015.

[103] Yang, M.; Zhang, L.; Feng, X. C.; Zhang, D. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision* Vol. 109, No. 3, 209–232, 2014.

[104] Xie, Y.; Zhang, W. S.; Li, C. H.; Lin, S. Y.; Qu, Y. Y.; Zhang, Y. H. Discriminative object tracking via sparse representation and online dictionary learning. *IEEE Transactions on Cybernetics* Vol. 44, No. 4, 539–553, 2014.

[105] Isard, M.; Blake, A. CONDENSATION—Conditional density propagation for visual tracking. *International Journal of Computer Vision* Vol. 29, No. 1, 5–28, 1998.

[106] Danelljan, M.; Hager, G.; Khan, F. S.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, 4310–4318, 2015.

[107] Li, F.; Tian, C.; Zuo, W. M.; Zhang, L.; Yang, M. H. Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4904–4913, 2018.

[108] Danelljan, M.; Bhat, G.; Khan, F. S.; Felsberg, M. ECO: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6931–6939, 2017.

[109] Lukežič, A.; Vojíř, T.; Čehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative correlation filter tracker with channel and spatial reliability. *International Journal of Computer Vision* Vol. 126, No. 7, 671–688, 2018.

[110] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representation, 2015.

[111] Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4293–4302, 2016.

[112] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[113] Li, C. L.; Liu, L.; Lu, A. D.; Ji, Q.; Tang, J. Challenge-aware RGBT tracking. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12367*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 222–237, 2020.

[114] Chen, X.; Yan, B.; Zhu, J. W.; Wang, D.; Yang, X. Y.; Lu, H. C. Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8122–8131, 2021.

[115] Yan, B.; Peng, H. W.; Fu, J. L.; Wang, D.; Lu, H. C. Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10428–10437, 2021.

[116] Hare, S.; Saffari, A.; Torr, P. H. S. Struck: Structured output tracking with kernels. In: Proceedings of the International Conference on Computer Vision, 263–270, 2011.

[117] Lukežič, A.; Zajc, L. Č; Vojíř, T.; Matas, J.; Kristan, M. Now you see me: Evaluating performance in long-term visual tracking. *arXiv preprint* arXiv:1804.07056, 2018.

[118] Wu, Y.; Lim, J.; Yang, M. H. Online object tracking: A benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2411–2418, 2013.

[119] Wu, Y.; Lim, J.; Yang, M. H. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 9, 1834–1848, 2015.

[120] Zhang, L. C.; Gonzalez-Garcia, A.; van de Weijer, J.; Danelljan, M.; Khan, F. S. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing* Vol. 28, No. 4, 1837–1850, 2019.

[121] Bendjebbour, A.; Delignon, Y.; Fouque, L.; Samson, V.; Pieczynski, W. Multisensor image segmentation using Dempster–Shafer fusion in Markov fields context. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 39, No. 8, 1789–1798, 2001.

[122] Xu, H. X.; Chua, T. S. Fusion of AV features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 2, No. 1, 44–67, 2006.

[123] Liu, H. X.; Simonyan, K.; Yang, Y. M. DARTS: Differentiable architecture search. In: Proceedings of the International Conference on Learning Representations, 2019.

[124] Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q. V. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8697–8710, 2018.

[125] Huang, C.; Lucey, S.; Ramanan, D. Learning policies for adaptive tracking with deep feature cascades. In: Proceedings of the IEEE International Conference on Computer Vision, 105–114, 2017.

[126] Guo, Q.; Feng, W.; Gao, R. J.; Liu, Y.; Wang, S. Exploring the effects of blur and deblurring to visual object tracking. *IEEE Transactions on Image Processing* Vol. 30, 1812–1824, 2021.

[127] Guo, Q.; Cheng, Z. Y.; Juefei-Xu, F.; Ma, L.; Xie, X. F.; Liu, Y.; Zhao, J. J. Learning to adversarially blur visual object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10819–10828, 2021.

Pengyu Zhang received his B.E. degree in information engineering from Chang'an University, China, in 2016. He is currently pursuing the Ph.D. degree with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology (DUT), Dalian, China. His research interests include visual object tracking and multiple modality fusion.

Dong Wang (Member, IEEE) received his B.E. degree in electronic information engineering and Ph.D. degree in signal and information processing from the Dalian University of Technology (DUT), Dalian, China, in 2008 and 2013, respectively. He is currently a professor with the School of Information and Communication Engineering, DUT. His research interests include object detection and tracking.

Huchuan Lu (Senior Member, IEEE) received his M.S. degree in signal and information processing and Ph.D. degree in system engineering from the Dalian University of Technology (DUT), China, in 1998 and 2008, respectively. He has been a faculty member (since 1998) and professor (since 2012) with the School of Information and Communication Engineering, DUT. His research interests include computer vision and pattern recognition. In recent years, he has focused on visual tracking and segmentation. He currently serves as an Associate Editor for *IEEE Transactions on Cybernetics* and *IEEE Transactions on Circuits and Systems for Video Technology*.