

Omni Survey for Multimodality Analysis in Visual Object Tracking

Zhangyong Tang, Tianyang Xu *Member, IEEE*, Xuefeng Zhu, Hui Li, Shaochuan Zhao, Tao Zhou, Chunyang Cheng, Xiao-Jun Wu*, and Josef Kittler *Life Member, IEEE*

Abstract—The development of smart cities has led to the generation of massive amounts of multi-modal data in the context of a range of tasks that enable a comprehensive monitoring of the smart city infrastructure and services. This paper surveys one of the most critical tasks, multi-modal visual object tracking (MMVOT), from the perspective of multimodality analysis. Generally, MMVOT differs from single-modal tracking in four key aspects, *i.e.*, data collection, modality alignment and annotation, model designing, and evaluation. Accordingly, we begin with an introduction to the relevant data modalities, laying the groundwork for their integration. This naturally leads to a discussion of challenges of multi-modal data collection, alignment, and annotation. Subsequently, existing MMVOT methods are categorised, based on different ways to deal with visible (RGB) and X modalities: programming the auxiliary X branch with replicated or non-replicated experimental configurations from the RGB branch. Here X can be thermal infrared (T), depth (D), event (E), near infrared (NIR), language (L), or sonar (S). The final part of the paper addresses evaluation and benchmarking. In summary, we undertake an omni survey of all aspects of multi-modal visual object tracking (VOT), covering six MMVOT tasks and featuring 338 references in total. In addition, we discuss the fundamental rhetorical question: Is multi-modal tracking always guaranteed to provide a superior solution to unimodal tracking with the help of information fusion, and if not, in what circumstances its application is beneficial. Furthermore, for the first time in this field, we analyse the distributions of the object categories in the existing MMVOT datasets, revealing their pronounced long-tail nature and a noticeable lack of animal categories when compared with RGB datasets. We conclude the discussions by highlighting promising directions for future research and provide a curated collection of up-to-date GitHub repositories for the MMVOT community with the link: <https://github.com/Zhangyong-Tang/Awesome-MultiModal-Visual-Object-Tracking>.

Index Terms—Multi-modal visual object tracking, omni survey, information fusion

I. INTRODUCTION

FACILITATING the development of smart cities is a forward-looking yet challenging objective in the field of

artificial intelligence. In these environments, intelligent and autonomous agents are expected to handle complex decision-making processes while ensuring public safety. However, this goal remains difficult to achieve due to the dynamic and unpredictable nature of real-world scenarios, particularly when agents operate with incomplete or fragmented perceptions of their surroundings. To address this, as shown in Fig. 1(a), researchers are increasingly focused on aggregating multi-modal information to construct a more reliable understanding of the environment, thereby enhancing the robustness of agents in all-day, all-weather applications [1], [2].

As one of the most critical competences enabling motion perception in intelligent systems, such as embodied robotics [12] and autonomous vehicles [13], visual object tracking (VOT) draws increasing attention and remains a prominent area of research in computer vision [14]–[17]. The goal of VOT is to maintain awareness (location and size) of an object, given its position in the first frame of an RGB sequence [16]. Analogously, multi-modal visual object tracking (MMVOT) tasks pursue the same objective but rely on multiple input signals (typically two streams), as illustrated in Fig. 1(b), to provide more robust perception in challenging conditions with the results exemplified in Fig. 2. In recent years, driven by the widespread availability of sensors that capture complementary information to RGB data, interest in MMVOT has surged, as clearly shown in Fig. 1(c).

The aim of this work is to reflect on this rapid development and to present a comprehensive survey of MMVOT tasks from the perspective focusing on multimodality. *Beyond the taxonomy of existing methods, it provides in-depth analyses of the physical principles underlying various data modalities, strategies for aligning and annotating multi-modal data, as well as the evaluation metrics and datasets used for performance assessment—thus offering a holistic overview of multimodality in MMVOT.* Furthermore, this work systematically investigates two significant, yet under-explored issues faced by the MMVOT community: (1) the relevance of applying multi-modal information fusion densely and indiscriminately in all scenarios, and (2) the scarcity of certain type of data, *e.g.* animals, in non-RGB streams, and the much more pronounced long-tail distribution of object categories compared with that in the RGB datasets. These issues are identified and discussed in this paper for the first time in this field. Finally, the paper concludes by outlining promising directions for future research. All relevant materials are made publicly available at our [homepage](#), which will be continuously updated to further benefit this community.

Z. Tang, T. Xu, X. Zhu, H. Li, C. Cheng, and X.-J. Wu are with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China. (First author: Z. Tang: zhangyong_tang_jnu@163.com, Corresponding author: X.-J. Wu, e-mail: wu_xiaojun@jiangnan.edu.cn)

S. Zhao is with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. (e-mail: shaochuan_zhao@cumt.edu.cn)

T. Zhou is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. (e-mail: : taozhou.ai@gmail.com)

Josef Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. (e-mail: j.kittler@surrey.ac.uk)

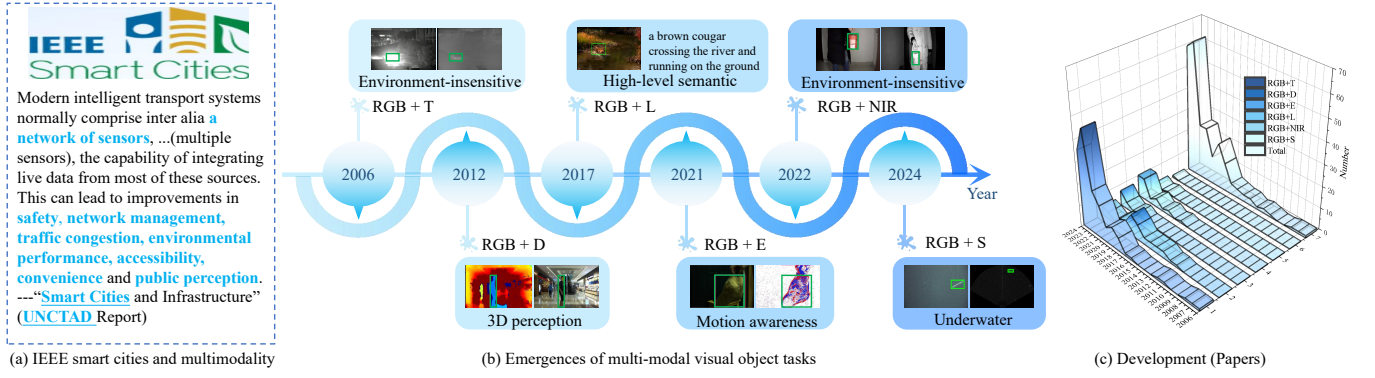


Fig. 1. (a) Multi-modal sensing is an essential element of smart cities. (b) The history of multi-modal visual object tracking. (c) The development of all MMVOT tasks in terms of papers published during 2006 to 2024.

TABLE I
A DIFFERENTIATION FROM THE EXISTING SURVEYS AND REPORTS. 'REF. NUM' MEANS THE NUMBER OF REFERENCES.

Type	Task	Citation	Year	Venue	Ref. Num	Involvement
Survey	RGB+T (2)	[3]	2020	INFFUS	142	Methods; Benchmarks; Future works
Survey	RGB+T (2)	[4]	2022	Arxiv	56	Methods; Benchmarks; Future works
Survey	RGB+T (2)	[5]	2023	TOMM	114	Methods; Benchmarks; Future works
Survey	RGB+T (2)	[6]	2024	TIM	158	Methods; Benchmarks; Future works
Survey	RGB+T (2)	[7]	2024	INFFUS	168	Methods; Benchmarks; Future works
Survey	RGB+D (2)	[8]	2022	Arxiv	75	Methods; Benchmarks; Future works
Survey	RGB+D (2)	[9]	2024	CADCG	86	Methods; Benchmarks; Future works
Survey	RGB+T/D (3)	[10]	2024	CVM	127	Methods; Benchmarks; Future works
Report	RGB+T/D/E/L (5)	[11]	2024	Arxiv	154	Methods; Benchmarks
Survey	RGB+T/D/E/L/S/NIR (6)	Ours	-	-	338	Multi-Modal Data Collection&Annotation&Alignment; Methods; Benchmarks; Future works

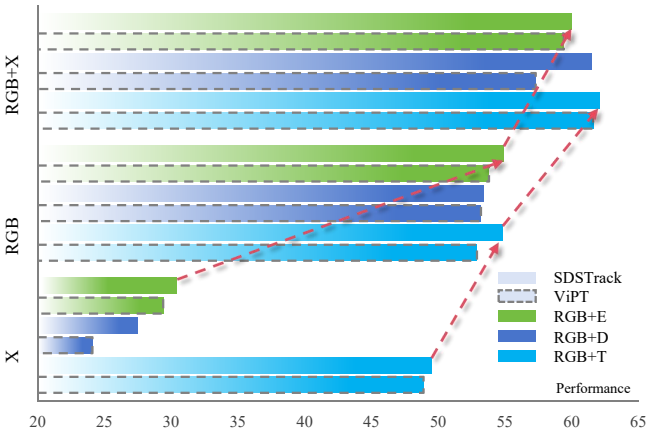


Fig. 2. An illustration of the evolution of the use of multi-modal data through experimental studies involving RGB+T, RGB+D, and RGB+E tracking tasks.

Related Surveys and Report: Given the rapid growth of interest in MMVOT tasks, we believe that an up-to-date and comprehensive survey from a holistic perspective is of vital importance to the community. Table I lists the existing surveys [3]–[10] and a report [11]. Most existing surveys focus on one or two MMVOT tasks, such as RGB+T, RGB+D, or RGB+D/T tracking. While the report in [11] covers a relatively wider range of tasks, it lacks a comprehensive and in-depth analysis. This limitation motivates our work to encompass a broader spectrum of MMVOT tasks and provide detailed discussions

and insights to further accelerate progress in the field.

Contributions: The contributions of our review of the landscape of the MMVOT developments can be summarised as follows:

- This work presents the first comprehensive survey encompassing RGB+T, RGB+D, RGB+E, RGB+L, RGB+NIR, and RGB+S tracking tasks. With a total number of 338 references, it offers readers a concise yet thorough understanding of multi-modal visual object tracking.
- Our survey of multi-modal VOT approaches is holistic, covering all aspects of MMVOT tasks, ranging from multi-modal data collection to multi-modal method evaluation.
- This work systematically discusses two significant yet previously overlooked issues: ① The need for indiscriminate and dense application of multi-modal fusion across all scenarios; ② The biased nature of object category distributions in MMVOT datasets, which exhibit clear long-tail characteristics that may hinder generalisation. Additionally, this work reveals, for the first time in the MMVOT community, the scarcity of animal data in non-RGB video streams.
- We also outline the warranted and promising directions for future MMVOT research, such as physics-inspired architectures and algorithms for modality customisation and developing metrics to quantify the quality of the respective modalities available for information fusion, aiming to uncover the hidden relationships between information fusion and tracking performance.

Scope: Before proceeding with the detailed analysis, we

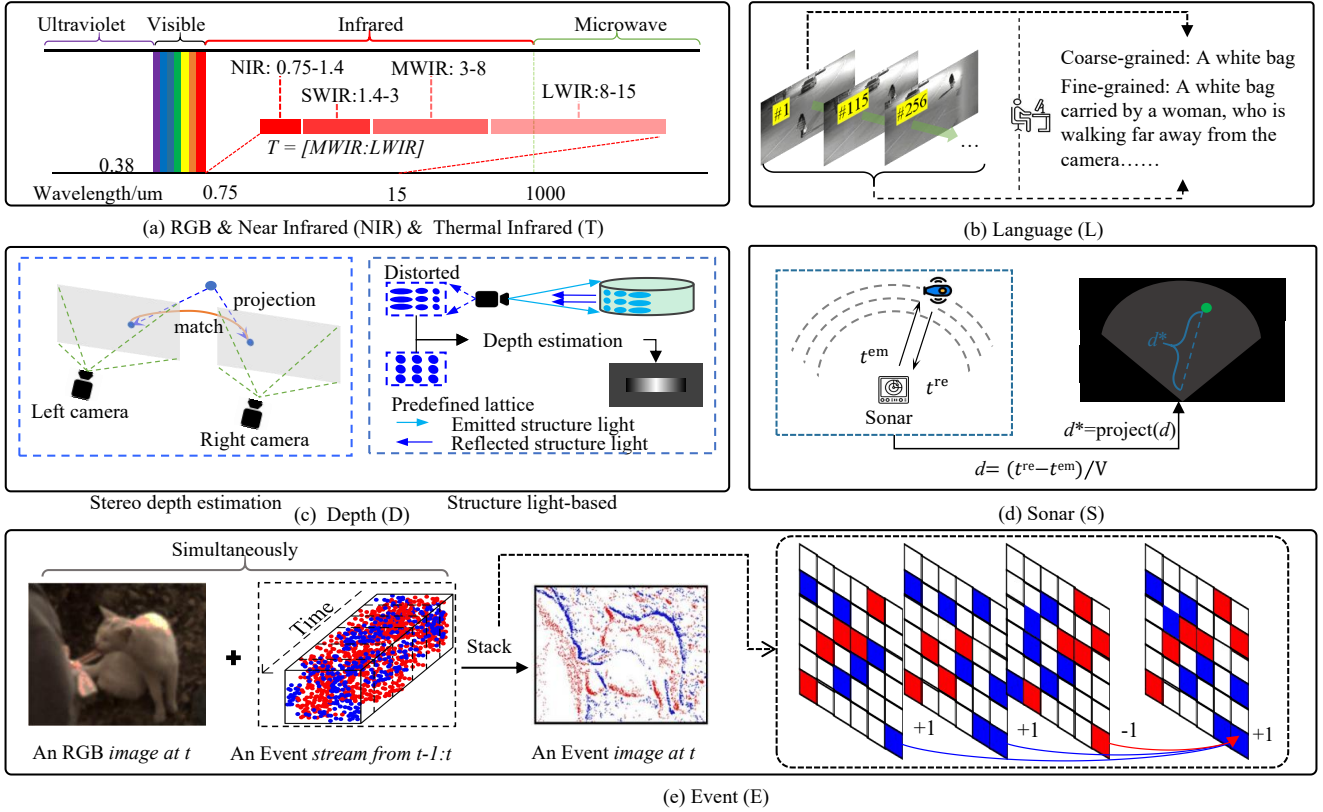


Fig. 3. Illustrations of the involved data modalities, including (a) RGB, T, NIR, (b) L, (c) D, (d) S, and (e) E modalities.

clarify the scope of this work: (1) The focus is on single object tracking - each sequence contains only one target trajectory; for multi-object tracking, please refer to [18]; (2) Generic object tracking - category-specific tracking methods, such as pedestrian tracking [19], are not considered; (3) 2-D object tracking - this survey focuses on bounding box-based tracking. For 3D object tracking, readers are referred to [20].

II. MULTI-MODAL DATA COLLECTION

To facilitate an in-depth understanding, first and foremost, **we begin by providing a physical introduction to the data modalities used for tracking, highlighting their unique and complementary characteristics in relation to RGB data, which is always ignored in related works.**

A. Data Modalities

Visible (RGB), Near Infrared (NIR), and Thermal Infrared (T): As depicted in Fig. 3(a), RGB data is captured by sensing the reflected visible spectrum with a wavelength between 0.38 and 0.75 μm ¹. Electromagnetic waves at different wavelengths exhibit distinct properties. For example, RGB data inherits the characteristics of the visible spectrum, making its imaging process (closely related to its quality) sensitive to environmental factors. This motivates the incorporation of other electromagnetic waves. Infrared radiation, emitted by objects themselves and less influenced by external conditions,

typically begins at wavelengths longer than the red portion of the visible spectrum (0.75–1000 μm ²). The infrared spectrum is further categorised into five parts: Near-Infrared (NIR, 0.75–1.4 μm), Short-Wavelength Infrared (SWIR, 1.4–3 μm), Mid-Wavelength Infrared (MWIR, 3–8 μm), Long-Wavelength Infrared (LWIR, 8–15 μm), and Far-Infrared (FIR, 15–1000 μm). Among these, MWIR and LWIR are collectively referred to as thermal infrared (T), which is particularly valuable for detecting thermal radiation emitted by objects themselves, thus being robust to environmental conditions.

Language (L): As shown in Fig. 3(b), unlike the conventional bounding box-based annotations, language offers a more natural and intuitive way for humans to describe objects, providing richer high-level semantic information [21], [22].

Sonar (S): In Fig. 3(d), sonar sensors operate by actively emitting ultrasound waves at a specific timestamp t^{em} . When these waves encounter an object, they are reflected back to the sensor and received at a later timestamp t^{re} . The time difference between emission and reception, $\Delta t = t^{\text{re}} - t^{\text{em}}$, is used to calculate the distance d between the sonar sensor and the object, given the known speed of sound, V , as follows: $d = (\Delta t * V)/2$. Finally, according to predefined rules, this distance is projected into the image space to create the sonar image: $d^* = \text{project}(d)$, which depicts distance information.

Depth (D): Depth information can be acquired through three main approaches: Time-of-Flight (ToF) methods, stereo depth estimation, and structured light-based techniques [23].

¹https://en.wikipedia.org/wiki/Visible_spectrum

²<https://en.wikipedia.org/wiki/Infrared>

The principle of ToF-based methods is similar to that of sonar data, relying on the temporal gap between emitting and receiving signals. For this reason, ToF is not illustrated in Fig. 3(c), while the other two methods are depicted. On the left side, stereo depth estimation uses two cameras that mimic the human visual system by capturing two slightly different views of the same scene. By comparing the disparity between these views, along with camera-specific parameters such as viewpoint and baseline distance between the cameras, the depth of objects can be estimated accurately. The right side of Fig. 3(c) illustrates the structure light-based method. In this approach, a camera projects a known light pattern (such as a grid or lattice) onto the object. The deformation of the pattern caused by the surface of an object is captured in the reflected light. By analysing this distortion, combined with internal camera parameters, the depth information of the object is derived. Additionally, stereo depth estimation is generally recommended for outdoor scenarios due to its robustness in such environments [23].

Event (E): Event cameras are biologically inspired by neurons, with values representing activation or inactivation. Unlike regular RGB cameras that capture intensity images in a frame-based manner with relatively high latency (10–20 ms), event cameras output a stream of asynchronous events. They record discrete changes in pixel intensity, indicated by positive (+1) or negative (−1) values, depending on whether the pixel intensity increases or decreases, respectively. This allows event cameras to perceive rapid motions with very low latency (1 us) [24]. As illustrated in Fig. 3(e), during the acquisition of the t -th RGB frame, an event stream spanning from $t - 1$ to t is obtained. This stream is considered the raw event data and can be represented as a voxel grid [25]. More commonly, the data is stacked to form the t -th event image [24], [26] by aggregating values at the same spatial location across all event slices, as shown in the right side of Fig.3(e).

B. Data Collection

After identifying the physical characteristics of each data modality, their complementary nature becomes apparent, encouraging researchers to explore multi-modal integration to develop more robust intelligent systems [1], [27]. Among the modalities discussed in this work, the L modality is unique in that it is derived from image data, with its description provided manually in early datasets [28], and more recently generated using large language models (LLMs) [29]. In contrast, other modalities are acquired through dedicated sensors. Typically, two cameras are mounted on the same platform to capture different modalities, such as the DLD-J18-161 for T and the DS-2ZCN3007 for RGB in [30], or an HD Zoom SeaCam underwater camera alongside an Oculus MD-Series sonar in [31]. While hand-crafted platforms integrating multiple sensors are widely used for data collection, they often demand a significant effort to ensure precise multi-modal alignment [32]. Alternatively, off-the-shelf systems such as the ZED 2i stereo camera³ simplify the process by providing factory-calibrated setups, thereby eliminating the need for complex

sensor calibration. However, these commercial solutions can be prohibitively expensive⁴, making hand-crafted platforms a more practical choice for many researchers in the community.

Typically, MMVOT tasks involve two separate data streams, with the exception of the RGB+NIR task. In this case, RGB and NIR data coexist within the same sequence. This is made possible by specialised cameras designed for RGB+NIR tracking, which can adaptively switch between RGB and NIR sensors in response to varying illumination conditions [33].

III. MULTI-MODAL DATA ALIGNMENT AND ANNOTATION

A. Data Alignment

Unlike single-modal data, multi-modal data typically requires alignment to address spatial discrepancies caused by different sensor placements. However, within the scope of this work, the RGB+NIR, RGB+L, and RGB+S datasets do not require such alignment, for the following reasons: RGB+NIR tracking involves a single video stream containing both modalities, captured by cameras that switch adaptively between RGB and NIR modes, making the data alignment-free [34]. RGB+L data is semantically aligned, as the linguistic descriptions are generated from RGB images. Moreover, the differing data formats (text vs. image) render traditional spatial alignment methods inapplicable [21]. RGB+S data, although both in image format, differs fundamentally in content. RGB captures visual appearance, while S encodes distance information. This semantic mismatch precludes meaningful spatial alignment [31].

In contrast, alignment is essential for the three most widely studied MMVOT tasks: RGB+T, RGB+D, and RGB+E tracking. For RGB+T datasets, the most common alignment strategies involve computing transformation matrices using either intrinsic and extrinsic camera parameters [30], [32], [35]–[37] or manually annotated key points [38]–[40]. For RGB+D datasets, alignment is typically handled internally by the hardware. Examples include stereo cameras like the ZED 2i [41], smartphones such as the iPhone [42], and depth sensors like the Kinect [23], [43]–[45] and ASUS Xtion [46]. RGB+E datasets are generally aligned within the sensor hardware, such as Dynamic Vision Sensors (DVS) [24], [47]. Notably, CRSOT [26] is an exception, as it is collected using an RGB camera and a higher-resolution event sensor without alignment, aiming to explore fusion strategies under unaligned conditions.

B. Data Annotation

Among the discussed modalities, L is unique in that it inherently serves as a form of high-level semantic annotation and therefore does not require traditional labelling. In contrast, all other modalities (RGB, T, D, E, NIR, S) rely on conventional annotation methods, typically involving the use of bounding boxes to identify target objects. Generally, two standard formats are employed to represent bounding boxes: $[x_{tl}, y_{tl}, x_{br}, y_{br}]$ [35] and $[x_{tl}, y_{tl}, w, h]$ [24], [30], [37], where subscripts “tl” and “br” denote the top-left and bottom-right corners, respectively. x and y represent the coordinates

³<https://www.stereolabs.com/en-cn/store/products/zed-2i>

⁴<https://store.dji.com/cn/product/zenmuse-h30t-and-dji-care-enterprise-basic?from=site-nav&vid=167271>

on the horizontal and vertical axes; w and h refer to the width and height of the bounding box. Thus, the first format defines a box using the coordinates of two opposing corners, while the second uses the top-left corner and the box dimensions.

In general, annotation requires experts familiar with the tracking process to ensure the precision and quality of the bounding boxes. However, this manual approach is both time-consuming and labour-intensive, which has motivated the adoption of semi-automatic annotation methods. In such methods, automatic tools, such as LabelMe⁵, are first used to generate coarse bounding boxes. These initial annotations are then manually refined to ensure accuracy. This approach is particularly efficient for simple or less dynamic videos, where the automatically generated boxes require minimal adjustment, thereby significantly reducing the overall annotation effort.

IV. MULTI-MODAL MODEL DESIGNING

As a rapidly emerging topic in the VOT community, MMVOT methods inherit many architectural designs and tracking strategies from RGB-only methods. Let $T_{\text{RGB}} : D_{\text{RGB}} \rightarrow \mathbb{R}^4$ denotes an RGB tracker, mapping the RGB input dataset D_{RGB} to bounding boxes in $\mathbb{R}^4$. Analogously, a multi-modal tracker can be defined as $T_{\text{MMVOT}} : D_{\text{MMVOT}} \rightarrow \mathbb{R}^4$, where D_{MMVOT} represents the multi-modal dataset and $\mathbb{R}^4$ remains the shared output space of the bounding boxes. This formulation highlights the strong methodological parallels between RGB-only VOT and MMVOT tasks. *Accordingly, taking the RGB branch as a reference, existing MMVOT methods can be classified into two main categories based on the design of the additional modality branch (X branch, $X \in [T, D, E, \text{NIR}, L, S]$): (1) Replicated Configuration - where the X branch mirrors the architecture of the RGB branch, and (2) Non-Replicated Configuration - where the X branch adopts a different architecture. Moreover, the methods with non-replicated configurations are specified with configurations customised or non-customised to the input of X branch. This taxonomy is conceptually illustrated in Fig. 4.*

As shown in Fig. 5, for clarity, the methods within each category are further grouped into three kinds: the methods primarily focusing on the straightforward information *Integration*, the methods learning *Boosted* feature representations, and the methods addressing *Real-world* issues. These sub-categories span a comprehensive design space for the MMVOT methods, offering a structured perspective for analysing and comparing different approaches.

A. X Branch with Replicated Configurations

The MMVOT methods that adopt the same experimental configurations for both the RGB and X branches fall into this category.

A.1 Early Stage Information Fusion. In the early stages, several attempts are made to simply aggregate the advantages of RGB and X modalities with features extracted by convolutional neural networks (CNNs) [48]–[55], deep Boltzmann machines [56], or hand-crafted descriptors such as histogram

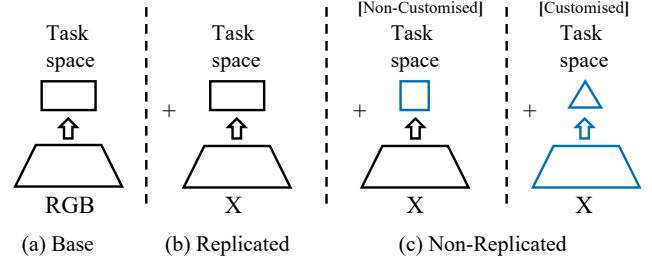


Fig. 4. A conceptual illustration of the taxonomy of the MMVOT methods.

of oriented gradients (HOG [57]) [58]. Although these methods demonstrate improved performance over single-modality approaches, they do not take advantage of the power of deep learning for information fusion.

A.2 Exploring Single-Level Fusion. Building on the rationale that the features at different network stages capture distinct semantic information [59], researchers have begun to investigate the benefits of fusion at different levels, including pixel-level fusion [60]–[62], feature-level fusion [31], [33], [34], [63]–[69], and decision-level fusion [70]–[77]. These investigations have collectively demonstrated the benefits drawn from each fusion level, thereby motivating further research into multi-level fusion strategies.

A.3 Exploring Multi-Level Fusion. Fusion of features from multiple levels is typically performed either simultaneously [78]–[82] or progressively [83]–[93]. In M2GCI [94], for instance, shallow features are fused simultaneously, while deep features are fused progressively. Here, simultaneous fusion refers to extracting features from different levels independently and fusing them afterward, whereas progressive fusion implies that the fused embeddings from earlier stages contribute to the feature extraction in subsequent layers. Among these, progressive fusion is more commonly adopted, as it reflects a coarse-to-fine reasoning paradigm, aligning with the hierarchical nature of visual representation [95].

However, the discriminative power of feature representations used in the aforementioned methods is not guaranteed, often resulting in suboptimal fusion. To address this, extensive efforts have been devoted to learning more robust and informative feature embeddings, which in turn enhance the effectiveness of the fusion process.

A.4 Learning Temporal, Frequency, and Spatial & Channel Enhancement. (1) Temporal: As a video understanding task, MMVOT inherently benefits from temporal cues, which are crucial for building stable and reliable intelligent systems. Temporal modelling in MMVOT is commonly achieved through several strategies: maintaining online templates [96]–[99], defining dynamic memory pools for temporal message propagation [100]–[106], computing cross-frame attention maps [107]–[110], embedding bounding box coordinates [111], and designing long-term recovery mechanisms [112]–[114]. (2) Frequency: Noting that the attention mechanism effectively functions as a low-pass filter [115], preserving high-frequency information is of critical concern. The solutions suggested to achieve high frequency preservation include high-pass convolutional layers [116], Wavelet Transforms [117],

⁵<https://github.com/wkentaro/labelme>

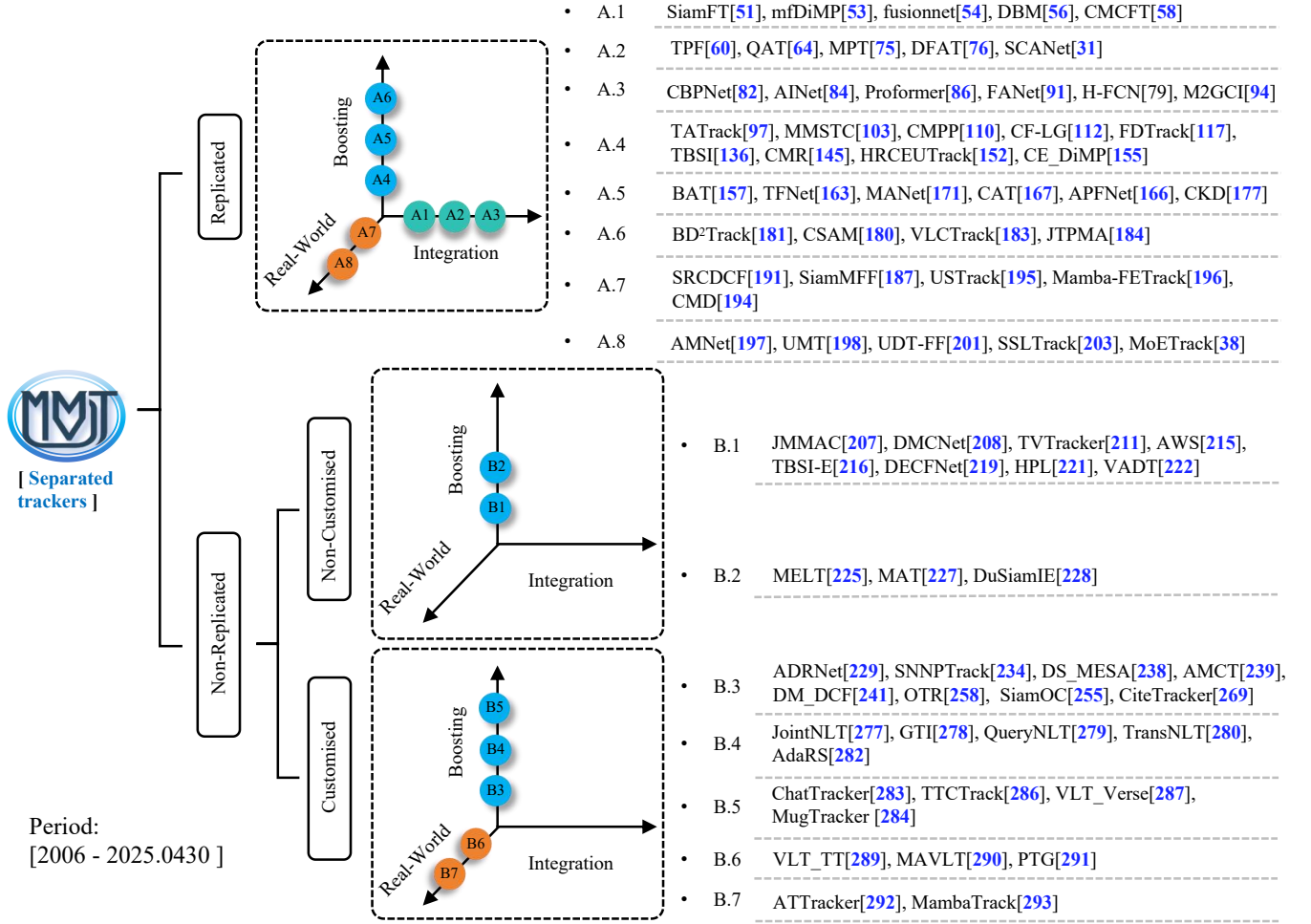


Fig. 5. The taxonomy of MMVOT methods according to the modality configurations and the corresponding representative trackers. A conceptual illustration of the taxonomy is provided in Fig. 4.

max pooling operations [118], and Fourier Transforms [119], [120]. (3) Spatial & Channel: Despite the advantages brought by temporal and frequency enhancements, most methods improve the representations along the spatial and channel dimensions by employing advanced attention mechanisms [121]–[137], adaptively distinguishing foreground from background [135], [138]–[151], facilitating high-quality multi-modal interactions [152], [153], mitigating spatial discrepancies across modalities [154], and swapping high-confidence channels within multi-modal embeddings [155].

A.5 Learning Upgraded Architectures and Losses. Beside feature enhancements, modifications of the network architecture and loss functions, such as the adaption of large pre-trained models [156]–[158], and the construction of multi-branch architectures tailored for different modalities [159]–[164] and for modality shared-specific challenging attributes [165]–[173], strive to achieve further performance uplift. Additionally, multi-modal contrastive loss functions have been employed to improve representation learning [174]–[178].

A.6 Learning Facilitated by Multi-Task Cooperations. Cooperative learning involving multiple tasks is instrumental to gaining further improvements by introducing auxiliary high-level guidance, in terms of multi-object tracking [179], [180],

image generation [181], image-text generation [182], [183], and segmentation [184]. Following the standard multi-task learning paradigm, the cooperative learning process can be optimised by introducing learnable weights to balance the multi-task losses [184], leading to a more harmonious and effective task integration.

Although promising performance has been achieved, several issues require further consideration for real-world deployment.

A.7 Computational Efficiency. Complex designs improve performance but often at the cost of lower efficiency, which remains a crucial and persistent challenge for the community. To accelerate MMVOT methods, researchers employ high-efficiency Siamese networks [185]–[189] or Discriminative Correlation Filter (DCF) frameworks [190], [191], lightweight and one-stream backbones [192]–[195] for feature extraction, as well as efficient fusion blocks such as Mamba [196].

A.8 Tracking in Adverse Scenarios. There is also considerable interest in the robustness of tracking under adverse conditions, which is critical for real-world deployment. This includes tracking with unaligned [197] and unpaired [198] data, handling multi-modal incomplete scenarios [38], [199], [200], and employing backbones trained via self-supervised learning [201]–[204].

B. X Branch with Non-Replicated Configurations

Mainstream methods often adopt identical configurations for both RGB and X branches. This approach overlooks the inherent physical differences between modalities, resulting in underutilisation of their complementary strengths. Motivated by this finding, many methods explore non-replicated configurations for the X branch, which can be either customised or non-customised. For non-customised methods, the focus is on two issues: effective leveraging of the benefits derived from the RGB modality and measure to enhance the representation of the X modality.

B.1 Exploiting the Strength of the RGB Modality. Compared with other data modalities, RGB data is known to excel thanks to its rich texture and colour information [30]. Detailed texture cues facilitate the estimation of object motion [205]–[207] and camera movement [208], [209], as well as the generation of linguistic descriptions [210], [211], which provide more appearance-independent semantic information. Colour information enables the use of Colour Name features [212] [213]–[215] and improves the perception of pixel intensity [216]. To leverage both, the texture and colour information, some methods dedicate special effort to the design of the RGB branch, focusing on the shallow layers, where such cues are abundant [217]–[220]. Another notable strength of the RGB modality is the widespread availability of visible-spectrum sensors, resulting in far larger datasets compared to other modalities. This abundance has enabled the development of large foundation models pre-trained solely on RGB data. Consequently, a prominent research direction in the MMVOT field is adapting these pre-trained models to multi-modal tracking tasks through techniques such as prompt-tuning [221], adapter-tuning [222], [223], and style transformation [224], aiming to maximise the benefits of these powerful models.

B.2 Exploiting Enhanced X Representations. Distinct from piggybacking on the RGB modality, a distinct trend is to enhance the representations derived for the X modality by training the X branch from scratch instead of fine-tuning [225], filtering out indiscriminate patterns through attention mechanisms [226], [227], and reusing X features to emphasize illumination-invariant patterns [228].

Recognising that non-customised configurations for the X branch achieve promising performance, researchers have taken a step further by designing customised X branches, tailored to the natural characteristics of each modality. The methods in this category feature modality-specific configurations and are therefore presented in the specialist sections dedicated to different modalities.

B.3 Building Modality-Customised Architectures. RGB+T: Based on its inherent imaging physics principles, RGB data is sensitive to illumination changes (ILL), whereas T data is not. Conversely, T data struggles in the presence of multiple objects with similar temperatures, which is called thermal cross-over (TC). Inspired by these insights, researchers have designed specialist branches, targeting these attributes: an ILL branch for RGB and a TC branch for T [229]–[231]. Each branch is trained separately using data annotated with the corresponding attribute.

RGB+E: E data can be regarded as binary images, where only the pixel positions with detected changes are activated, making the image much sparser than other modalities and functionally similar to neurons. Inspired by this sparsity, Pooler [232] leverages pooling operations to process event data effectively. Additionally, the methods based on Spiking Neural Networks (SNN) [233], such as [234]–[237], draw motivation from the biological characteristics of event data. Furthermore, most existing methods stack continuous event slices into a single event frame, as illustrated in Fig. 3(e), which tends to neglect fine-grained motion information. To overcome this limitation, the event stream can be divided into multiple groups, producing several sub-frames [238] that better preserve temporal details.

RGB+D: D data augments the RGB object description by enriching it from a unique distance perspective, providing consistent or smoothly varying depth values. Leveraging this property, depth information is used for position correction [239], [240], foreground/background segmentation [241], [242], and occlusion handling and indication [241], [243]–[255]. They commonly exploit the consistency of the depth values in the temporal domain based on the assumption only slight changes occur between adjacent frames or determine whether there are multiple peaks of interest in depth regions [256], [257]. Moreover, depth clues also provide 3-D perception, enabling the transformation of 2D images into 3D space, which allows for more precise object localisation and tracking [258]–[262].

RGB+L: Compared with other modalities, the L modality is not image-based, highlighting its unique characteristics and prompting the design of specialised architectures. Text encoders, which are tailored for processing linguistic inputs, are exclusive components of RGB+L tracking systems and do not appear in other multi-modal tracking tasks. Beyond this distinction, the development of RGB+L tracking largely parallels that of other modalities, with considerable efforts devoted to spatio-temporal enhancement [263]–[275].

B.4 Learning Facilitated by Multi-Task Cooperation. RGB+L: Due the inherent similarity of RGB+L tracking and visual grounding tasks, aiming to localise objects based on both visual-linguistic vs. linguistic inputs, numerous methods have explored cooperative learning between these two domains [276]–[281]. In addition, AdaRS [282] introduces a novel approach by treating linguistic descriptions as information queries. It retrieves candidate regions from RGB images based on high semantic similarity, thereby facilitating an effective integration of RGB and L modalities.

B.5 Learning Refined X Representations. RGB+L: As introduced in Sec. II-A, linguistic annotations in RGB+L tracking typically describe the target object in the entire video sequence, specific video segments, or only the initial frame, rather than providing frame-level descriptions. This coarse granularity implies that the annotations may not fully capture the dynamics of object appearance across time, and inaccuracies in the textual descriptions further exacerbate the issue. To address these limitations, recent methods have explored refining text embeddings through updating textual inputs during tracking [283], [284], adaptively selecting the most discriminative linguistic patterns [285], [286], and main-

TABLE II
AN OVERVIEW OF EXISTING UNIFIED MMVOT TRACKERS. PLEASE FIND THE DEFINITION OF "UNI-A" AND "UNI-B" IN SEC.IV-C.

Unification	Method	Venue	Backbone	LasHeR	DepthTrack	VisEvent	TNL2K	RGBS50	Descriptions
Uni-A	ProTrack [294]	ACM MM'22	ResNet [295]	42.1	57.8	47.4	-	-	Pioneering work; Prompt-tuning
	ViPT [296]	CVPR'23	ViT-B	52.5	59.4	59.2	-	-	Milestone; Prompt-tuning
	FDAFT [297]	PRCV'23	ViT-B [298]	55.3	62.0	-	-	-	-
	MINet [299]	IVC'24	ViT-B	52.9	60.4	59.4	-	-	-
	MixRGBX [300]	NeuroCom'24	MixFormer [301]	53.6	60.1	60.2	-	-	-
	AMATrack [302]	TIM'24	ViT-B	-	61.8	-	-	-	Generative Model; Tested on RGBT234
	KSTrack [303]	TCSVT'24	ResNet	-	58.7	57.2	-	-	Modality incomplete
	GMMT [304]	AAAI'24	ViT-B	56.6	-	-	-	-	Also tested on RGBD1K
	OneTracker [305]	CVPR'24	ViT-B	53.8	60.7	58.0	-	-	-
	SDSTrack [306]	CVPR'24	ViT-B	53.1	61.4	59.7	-	-	-
	SeqTrackV2 [307]	Arxiv'24	ViT-L	61.0	62.3	63.4	62.4	-	-
	STTrack [308]	AAAI'25	ViT-B	60.3	63.3	61.9	-	-	-
	CMDTrack [309]	TPAMI'25	ViT-B	56.6	59.8	61.3	-	-	Efficiency; Knowledge distillation
	LightFC-X [310]	Arxiv'25	ViT-T	50.7	53.8	53.2	-	58.6/39.5	Efficiency
Uni-B	EMTrack [311]	TCSVT'24	ViT-B	53.3	58.3	58.4	-	-	Efficiency; Knowledge distillation
	Un-Track [312]	CVPR'24	ViT-B	51.3	61.0	58.9	-	-	Still require task prior
	XTrack [313]	Arxiv'24	ViT-B	52.5	59.7	59.1	-	-	Mixture-of-expert
	M ³ Track [314]	SPL'25	ViT-B	51.1	59.7	59.2	-	-	Meta-learning [315]
	APTrack [316]	TAI'25	ViT-B	58.9	62.1	61.8	-	-	-
	UASTrack [317]	Arxiv'25	ViT-B	57.0	62.8	61.0	-	-	Modality-specific adapters
	SUTrack [318]	AAAI'25	Fast_iTPN-L [319]	61.9	66.4	63.8	67.9	-	Training with more data (RGB data)

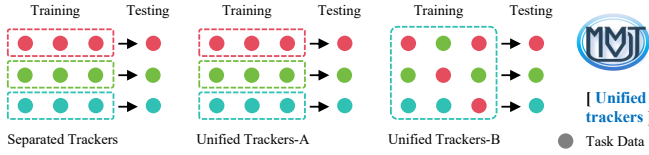


Fig. 6. Training and testing paradigms of existing unified trackers.

taining fine-grained attribute-level descriptions [287], [288].

B.6 Tracking with High Efficiency. RGB+L: Efficiency remains a long-standing and critical topic in the tracking community. Accordingly, several RGB+L tracking methods have focused primarily on improving efficiency. Specifically, this is achieved through the adoption of lightweight backbone networks [289], the use of more efficient fusion modules such as Mamba [290], and strategies that compress input information into fewer tokens [291].

B.7 Tracking in Adverse Scenarios. RGB+L: Robustness in adverse scenarios is also of significant concern. Hence, [292] explores a semi-supervised learning paradigm to mitigate the challenges posed by limited annotations and [293] focuses on enhancing the tracking performance in nighttime conditions, specifically within Unmanned Aerial Vehicle (UAV)-based settings.

C. Unified Multi-Modal Trackers

The methods shown in Fig.5 represent separate approaches, which are trained and evaluated on a single type of MMVOT task. However, the growing complexity and uncertainty in smart city environments demand more versatile solutions that are applicable to more than one task, prompting the study of pairwise multi-modal tracking tasks concurrently. This has led to the emergence of a new research direction, which focuses on the development of unified methods, capable of supporting multiple MMVOT tasks simultaneously, as illustrated in Fig.6. These unified methods can be categorised into two groups based on their training paradigms: Uni-A and Uni-B. Methods in Uni-A adopt a training and testing pipeline similar to developing separate methods, where each task is

handled independently, with solely a shared network architecture maintained and the parameters retrained. In contrast, Uni-B methods train on a mixture of data from all the tasks jointly, resulting in unified models that generalise across modalities and tasks. Detailed information on these unified methods, including publication source, year, backbone, performance, and brief descriptions, is provided in Table. II. This research direction has gained momentum since its inception in 2022, with a significant surge in interest for Uni-B methods, all of which have been published in 2024 and 2025.

It should be noted that all unified methods can also be classified following the same taxonomy used for separate methods, with the majority falling under the replicated and non-customised categories. However, our proposed taxonomy for unified methods offers a more granular perspective, clearly delineating the boundary between two distinct types of unification and providing deeper insight into their design philosophies.

D. Loss Function

The MMVOT and VOT tasks share the same objective, i.e. predicting the scale and location of a specified object. Hence, MMVOT trackers adopt the same loss functions as those used in RGB-only methods [320]. In view of this familiarity, due to the space limitations, the details of the loss functions are relegated to the supplementary material.

V. EVALUATION METHODOLOGY

In the following, we introduce the multi-modal datasets, evaluation metrics, the performance of existing methods, and the challenging attributes for detailed analysis.

A. Datasets

Table. III provides the key statistics for all MMVOT datasets, along with URLs to their homepages for easy access. From the table, it is evident that the number of datasets has increased significantly in recent years, particularly since 2020, reflecting the rapid development of the MMVOT field. Due

TABLE III
THE STATISTICS OF REAL WORLD MULTI-MODAL VISUAL OBJECTING TRACKING DATASETS.

Task	Dataset	Venue	Video Num.	Total Frames	Avg. Frames	Aligned	Annotation	Description
RGB+T	OTCBVS [39]	CVIU'06	6	8.5K	1424	✓	BB	the first RGB+T dataset
	LITIV [40]	CVIU'11	9	6.3K	702	✓	BB	pioneering dataset
	GTOT [35]	TIP'16	50	7.8K	157	✓	BB	the first dataset with certain scale
	RGBT210 [36]	ACM MM'17	210	104.7K	498	✓	BB	larger size
	RGBT234 [37]	PR'19	234	116.7K	498	✓	BB	RGBT210 + videos in some special scenarios
	LasHeR [30]	TIP'21	1224	734.8K	600	✓	BB	contains the first large-scale training set
	VTUAV [32]	CVPR'22	500	1.7M	3329	✓	BB+Mask	captured by UAVs
	MV-RGBT [38]	Arxiv'24	122	89.9K	737	✓	BB	considerations on modality validity
RGB+D	BoBoT_D [321]	-	5	-	-	-	-	unavailable, mentioned in [239]
	PTB [43]	ICCV'13	100	21.5K	215	✓	BB	pioneering dataset
	STC [46]	TCYB'17	36	9.0K	250	✓	BB	outdoor videos
	CDTB [45]	ICCV'19	80	102.0K	1274	✓	BB	more diverse
	DepthTrack [44]	ICCV'21	200	294.6K	1473	✓	BB	the first training set
	RGBD1K [41]	AAAI'23	1050	2.5M	2384	✓	BB	the first large-scale training set
	ARKitTrack [42]	CVPR'23	300	229.7K	765	✓	BB+Mask	collected by mobile phone
	D ² Cube [23]	CVPR'23	1000	1.0M	1030	✓	BB	captured by UAVs
RGB+E	EED [322]	IROS'18	7	234.0	33	✗	BB	the first dataset
	FE108 [25]	ICCV'21	108	208.7K	1932	✓	BB	the first dataset with certain scale
	VisEvent [24]	TCYB'24	820	371.1K	452	✓	BB	the first large-scale training set
	COESOT [323]	Arxiv'22	1354	478.7K	353	✓	BB	maximum number of videos
	FE240hz [224]	CVPR'23	110	143.0K	1300	✓	BB	an extension of FE108
	FELT [324]	Arxiv'24	742	1.6M	2148	✓	BB	long-term tracking
	CRSOT [26]	Arxiv'24	1030	305.0K	296	✗	BB	unaligned multi-modal data
	FE141 [325]	IJCV'24	141	25.1K	178	✓	BB	an extension of FE108
RGB+L	OTB99-L [326]	CVPR'17	100	59.0K	590	-	BB	the first dataset
	LaSOT [28]	CVPR'19	1400	3.5M	2506	-	BB	the first large-scale training set
	LaSOT_EXT [327]	IJCV'21	1550	3.9M	2502	-	BB	an extension of LaSOT
	TNL2K [21]	CVPR'21	2000	1.2M	622	-	BB	linguistic descriptions for the whole video
	MGIT [29]	NIPS'23	150	2.0M	13333	-	BB	3 multi-granularity text descriptions for long-term videos
	VastTrack [22]	NIPS'24	50610	4.2M	83	-	BB	the most object categories and videos
	VLT-MI [328]	Arxiv'24	3619	6.6M	1824	-	BB	multiple multi-modal interactions
	WebUOT-1M [329]	Arxiv'24	1500	1.1M	733	-	BB	underwater tracking
	DTVLT [330]	Arxiv'24	13134	8.0M	611	-	BB	4 multi-granularity text descriptions
	CMOTB-easy [33]	AAAI'22	644	478.0K	742	-	BB	the first dataset
RGB+NIR	CMOTB [34]	TNNLS'24	1000	799.0K	799	-	BB	an extension of CMOTB-easy
RGB+S	RGBSS0 [31]	TCSVT'24	50	43.7K	874	✗	BB	the first dataset
RGB+Multi	WebUAV-3M [331]	TPAMI'23	4500	3.2M	710	✓	BB	RGB+Language+Audio
	UniMod1K [332]	IJCV'24	1050	2.5M	2384	✓	BB	RGB+Depth+Language
	QuadTrack600 [333]	Arxiv'25	600	348.7K	581	✓	BB	RGB+Thermal++Event+Language, unpublished

to page limitations, only the most important information is included in Table. III. More detailed descriptions are available in the supplementary material.

B. Evaluation Metrics

To make this work self-contained, we introduce the widely used metrics, including precision rate (PR) [30], success rate (SR) [30], normalized precision rate (NPR) [334], precision (Pr) [45], recall (Re) [45], F-score [45], and the VOT protocol [335]. Due to page limitations, only the theoretical concepts are presented here, while detailed mathematical formulations are provided in the supplementary material.

PR and NPR: PR measures the percentage of accurately tracked frames, where the Euclidean distance between the centres of the ground truth and predicted bounding boxes falls below a predefined threshold. However, PR is sensitive to object size and image resolution, which can result in unfair evaluations for objects of different scales [334]. To address this, NPR is introduced as a normalised version of PR, where the Euclidean distance is divided by the diagonal length of the ground truth bounding box to enable a scale-invariant evaluation.

SR: It is a metric similar to PR, measuring the percentage of the successfully tracked frames, where the Intersection over Union (IoU) between the ground truth and predicted bounding boxes exceeds a predefined threshold (commonly set to zero).

Pr, Re, and F-score: They are long-term metrics borrowed from [45]. Pr and Re measure the average accuracy when the predicted and ground truth bounding boxes are present, respectively. The F-score combines Pr and Re to provide a comprehensive metric for ranking the tracker performance.

VOT protocol: Accuracy (A), robustness (R), and expected average overlap (EAO) are the metrics employed in the VOT protocol. Among them, A and R correspond closely to precision rate (PR) and success rate (SR), respectively, and EAO serves as a comprehensive metric that combines both A and R, playing a role similar to the F-score.

C. Benchmarking Results

Due to page limitations, Fig. 8 includes benchmarking results of four representative benchmarks, each displaying the performance of only the top 15 tracking methods. These benchmarks are LasHeR [30] (RGB+T), DepthTrack [44] (RGB+D), VisEvent [24] (RGB+E), and TNL2K [21] (RGB+L). As to the other two newly-emerged tasks, RGB+NIR and RGB+S, the maturity of research is very limited, which constrains in-depth discussion. Additional details, including the performance of all methods across all benchmarks and tracking efficiency, are available on the project [homepage](#).

Specially, all unified trackers are summarised in Table. II, providing a direct and comprehensive comparison of algo-

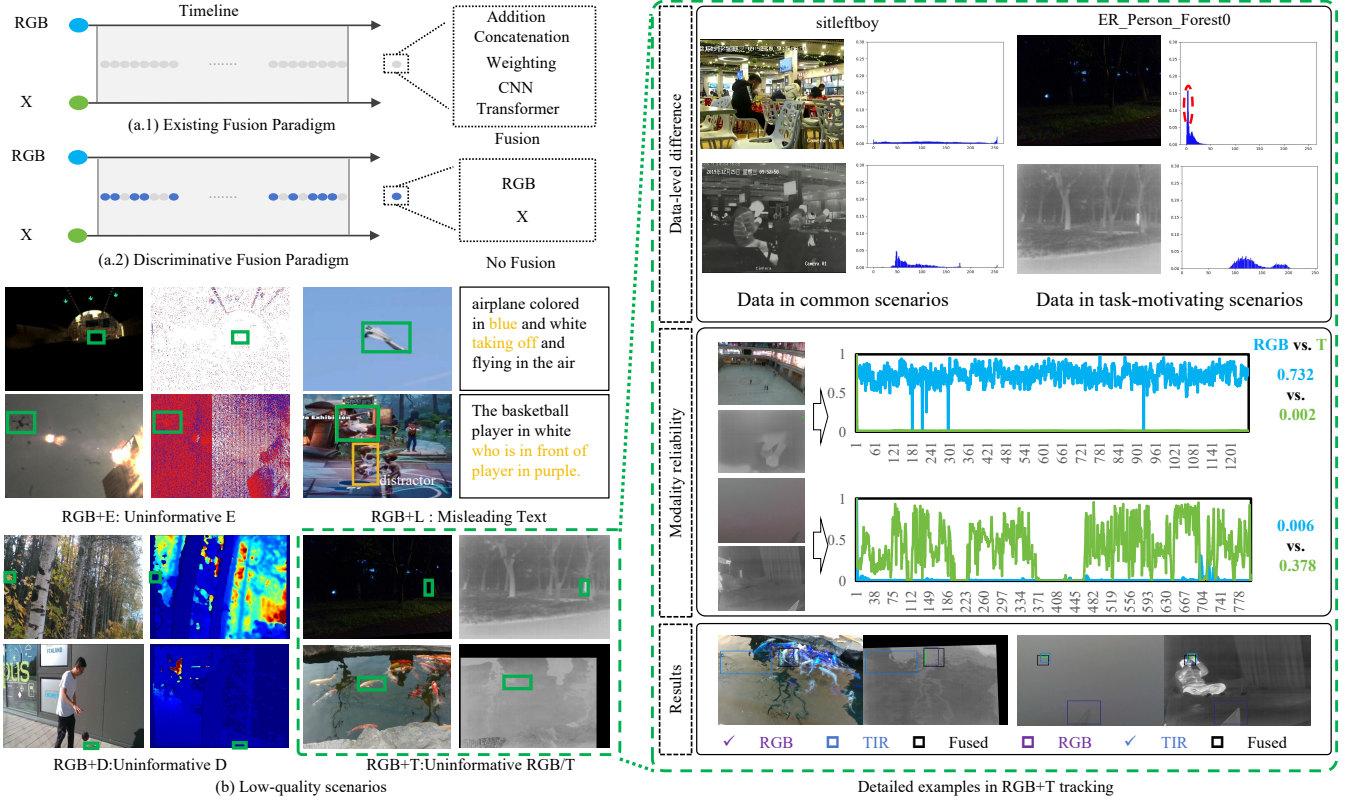


Fig. 7. (a) The differences between the existing and discriminative fusion paradigms. (b) An illustration of the scenarios where one modality suffers from low quality, exemplified by an example in RGB+T tracking.

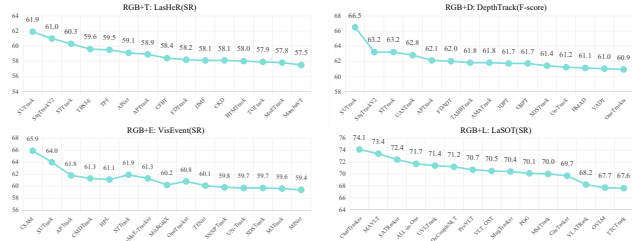


Fig. 8. The results on the most common benchmarks for each tracking task.

rithms targeting this emerging research area. Among these methods, SUTrack [318] achieves leading performance across all evaluated benchmarks, which can be attributed to four key factors: (1) a more powerful backbone, (2) task token recognition, (3) learnable background/foreground enhancement, and (4) the inclusion of additional RGB data during training. Among these, the utilisation of RGB data in the training phase is the most distinctive feature compared with other unified methods. This strategy is formulated as follows:

$$\begin{aligned} T_{\text{SUTrack}} &: (D_{\text{RGB}}, D_{\text{MMVOT}}, \theta_{\text{RGB}}) \rightarrow \mathbb{R}^4 \\ T_{\text{Others}} &: (D_{\text{MMVOT}}, \theta_{\text{RGB}}) \rightarrow \mathbb{R}^4 \end{aligned} \quad (1)$$

where T_{SUTrack} and T_{Others} denote SUTrack and other unified trackers, respectively. θ_{RGB} represents a model pre-trained on RGB data. As shown in Eqn. 1, SUTrack achieves unification by fine-tuning the RGB pre-trained model, while also incorporating the original RGB data during training. This

strategy aligns with the concept of *Replay* [336] in the field of continual learning, which preserves the performance achieved on previous tasks, while acquiring knowledge for new ones. By leveraging replay, SUTrack is able to retain the discriminative power of the RGB pre-trained model. This is particularly beneficial for tracking, as RGB-only datasets typically offer higher-quality data than most MMVOT datasets.

D. Challenging Attributes

To thoroughly analyse the strengths and limitations of various methods, all datasets listed in Table.III are annotated with multiple challenging attributes. In total, 61 distinct attributes are identified and categorised into task-shared and task-specific groups, based on their presence across tasks. However, limited by pages, detailed descriptions of these attributes are provided in the supplementary material.

VI. DISCUSSIONS

Discussion-A: Should multi-modal fusion be applied routinely or selectively? Current research largely focuses on designing increasingly complex fusion models to boost benchmark performance, often overlooking a more fundamental question, the benefit of fusion in particular circumstances itself. As illustrated in Fig. 7(a.1), most existing methods apply fusion in a dense manner, operating under the assumption that both modalities consistently offer high-quality information beneficial for accurate object localisation. However, this



Fig. 9. The differences in the distributions of the RGB, RGB+D, RGB+T, and RGB+S datasets. Note, sequences relating to other MMVOT tasks are not named according to specific object categories and are therefore unsuitable for this visualisation, except for RGB+L datasets, which share the same sequences as the RGB datasets.

assumption can be violated in both theoretical and practical contexts. Theoretically, task definitions like RGB+T tracking are motivated by specific scenarios such as nighttime, where the RGB modality suffers from low visibility and the T data becomes crucial [30]. We refer to such contexts as task-specific scenarios. Practically, as shown in Fig. 7(b), low-quality inputs from one or more modalities can be observed in datasets across a range of tasks, including RGB+E, RGB+L, RGB+D, and RGB+T tracking. This confirms that degraded modalities are not uncommon in real-world data. These theoretical motivations and empirical findings raise a critical question: Should multi-modal fusion be applied discriminately, such that low-quality modalities are excluded from the fusion process? A conceptual illustration of this discriminate fusion paradigm is shown in Fig. 7(a.2), where fusion is selectively activated based on data quality.

To provide a more intuitive understanding, the right part of Fig. 7 presents illustrative examples from RGB+T tracking. First, histograms depicting data distributions from both common and task-specific scenarios are shown, where one modality exhibits insufficient quality to support tracking effectively, highlighting the importance of taking the variability of data quality across different conditions into consideration. The subsequent quantitative analyses compare the performance of using RGB-only versus T-only inputs. The results indicate that using the less informative modality can yield a performance comparable to random noise (RGB-only vs. T-only: 0.732 vs. 0.002 and 0.006 vs. 0.378 on two cases). This finding suggests that fusion may be unhelpful or even harmful when one modality lacks any informative content. Subsequently, the visualisations of the tracking results using RGB, T, and RGB+T (fused) inputs demonstrate that fusion does not always lead to improvements compared with the deployment of a single high-quality modality. To sum up, these detailed examples strongly support the argument that multi-modal fusion should be applied discriminately. Routine fusion may inadvertently incorporate misleading or noisy information from the weaker

modality, thereby degrading the overall tracking performance, especially in task-specific scenarios where the quality of the respective modalities is highly imbalanced [38].

Difficulty: To address this, a data quality dependent fusion block must demonstrate robust performance across diverse scenarios. However, in some task-specific scenarios, no corresponding training datasets currently exist for all the tasks. For instance, only a test set is provided for RGB+T tracking [38]. This precludes the training of a discriminate fusion block.

Discussion-B: Are existing multi-modal datasets capable of enabling sufficient generalisation? In practical applications, the categories of tracked objects can be highly diverse, such as persons, vehicles, bags, and animals, depending on the scenarios. This diversity imposes a requirement on the MMVOT training datasets to cover a wide range of object categories in order to enhance the generalisation capability. To investigate this, Fig. 9 presents a qualitative analysis of several MMVOT datasets alongside LaSOT [28], a representative RGB-only tracking benchmark. The figure visualises the distribution and frequency of object categories within each dataset, offering insights into their potential for achieving good generalisation.

Compared with LaSOT, it is evident that object categories in existing MMVOT datasets are highly concentrated, often exhibiting pronounced long-tail distributions. Notably, we are not claiming that long-tail distributions are inherently undesirable. In fact, they are more reflective of the natural data distributions [22]. However, such imbalance can lead to biased model capabilities. In many cases, a MMVOT dataset contains a single dominant category. The models trained on such a dataset would perform significantly better on that category than on others. This contradicts the fundamental objective of visual object tracking: to robustly follow any specified object, regardless of its category, throughout a sequence.

In addition, a closer inspection reveals that animals are significantly under-represented in current MMVOT datasets, with nearly no presence in Fig. 9. This is noteworthy because

animal tracking constitutes a common and practical application scenario, where multi-modal data is often leveraged to enhance robustness [337]. The lack of multi-modal videos featuring animals therefore hinders progress in developing and evaluating trackers for such applications. In summary, existing multi-modal datasets exhibit suboptimal object category distributions, with the notable absence of animal data posing a particular limitation.

VII. FUTURE WORK

The preceding systematic review has comprehensively characterised the key components of MMVOT development, including multi-modal data collection (Sec.II), data alignment and annotation (Sec.III), model development (Sec.IV), and evaluation methodology (Sec.V), thereby establishing a holistic overview of the MMVOT field. Building upon these, we identify open challenges and research issues within each section, offering valuable suggestions for future investigations.

A. Multi-Modal Data Collection

- *Collecting Datasets with More Modalities:* Building unified trackers is currently a hot topic in the community. The models are typically trained on a mixture of multi-modal data from various tasks but are evaluated on benchmarks, each relating to a single type of task-specific data, leading to an inconsistency between training and testing procedures [296], [306], [314], [317]. To address this discrepancy, it would be valuable to construct a benchmark that incorporates multiple data modalities within a unified test set, thereby enabling more consistent and comprehensive evaluations of unified trackers.

- *Collecting Datasets in Task-Specific Scenarios:* For convenience, existing datasets are mainly collected in common scenarios, like sunny days, where both modalities present sufficient quality for supporting tracking tasks. However, the emergence of MMVOT tasks is typically driven by real-world scenarios where RGB data quality is compromised [21], [24], [31], [33], [38], [41]. This highlights a fundamental mismatch: current datasets often fail to represent the very conditions that motivate the use of multi-modal tracking methods. Therefore, it is imperative to construct new datasets captured in task-specific scenarios that, by design, address the data gap between controlled research settings and practical applications.

- *Collecting More Animal Data for MMVOT Tasks:* As discussed in Sec. VI, animal tracking in the wild represents a typical and meaningful application scenario, where the cooperation of multiple modalities is crucial for robust perception. However, this scenario is severely under-represented in existing MMVOT datasets, with a noticeable scarcity of multi-modal animal videos. To foster progress in this important direction, it is essential to collect and curate more multi-modal datasets specifically targeting animal tracking in diverse and challenging environments.

B. Multi-Modal Data Annotation

- MMVOT tasks are generally regarded as downstream extensions of the traditional VOT task. Consequently, many

configurations, including annotation protocols, are directly inherited. However, this practice does not always align with the physical characteristics of the additional modalities. For instance, T data reflects endogenous thermal radiation, where object temperature gradually decreases from the interior to the exterior until reaching equilibrium with the environment. This leads to inherently blurred object boundaries in thermal imagery, introducing uncertainty and ambiguity in bounding box annotations. Therefore, it is pertinent to explore modality-specific annotation strategies that better reflect the characteristics of each modality, aiming to leverage their complementary strengths, while mitigating the impact of noisy or ambiguous cues.

C. Multi-Modal Model Design

- *Discriminate Fusion Paradigm:* The need for multi-modal fusion is often justified by task-specific scenarios. For instance, under extreme illumination conditions, RGB images may become unusable, while T data remains unaffected. However, the prevailing fusion paradigm in MMVOT applies dense fusion at every frame, regardless of the quality of each modality. This observation highlights the limitations of indiscriminate fusion and underscore the importance of developing more selective, data quality dependent fusion paradigms, which can autonomously decide when and how to fuse modalities, improving robustness and reliability across both typical and adverse scenarios.

- *Physics-Induced Architectures:* As introduced in Fig. II-A, different modalities exhibit diverse physical characteristics. However, due to the predominance of RGB pre-trained models, most auxiliary modalities (except L) are processed using similar operations, such as converting data into a 3-channel format to fit the input requirements of RGB models, thereby overlooking their unique physical properties. To fully exploit the potential of these auxiliary modalities, it would be worth investigating the merits of developing physics-informed neural networks [338] that explicitly incorporate their inherent physical characteristics, enhancing tracking performance, efficiency, and interpretability simultaneously.

- *Multi-Modal LLMs in MMVOT:* Despite the significant breakthroughs in multi-modal LLMs [339], their application within the MMVOT domain is mainly limited to RGB+L tracking, remaining underexplored to a large extent. Given their broader knowledge topologies and reasoning capabilities, future research on integrating multi-modal LLMs into MMVOT tasks holds great potential, paving the way toward more reliable and general artificial intelligence in smart cities.

D. Multi-Model Method Evaluation

- Existing metrics for MMVOT tasks focus on evaluating tracking performance. Although this is the ultimate goal of tracking tasks, the assessment of multi-modal information fusion is largely overlooked and remains a black art. Many advanced methods claim improvements by more effective aggregation of multi-modal data. However, the improved performance only indicates that multi-modal information is better-used rather than better-fused. As illustrated in the right part

of Fig. 7, superior results can be achieved by relying on a single modality rather than their fusion. Therefore, developing metrics, which reflect the relationship between fusion and tracking performance better, is urgently required. Measuring the quality and effectiveness of multi-modal fusion would offer a valuable insight that could be harnessed to improve tracking.

VIII. CONCLUSION

This work presents the first omni-survey on multi-modal visual object tracking (MMVOT) from the perspective of multi-modal data properties and analysis. Our comprehensive discussion begins with an introduction to the physical characteristics of diverse modalities, revealing their complementary nature and providing a foundation for their deeper analysis. Building on this, the survey, encompassing 338 papers, covers all aspects of constructing a model to perform a multi-modal tracking task. The topics include data collection, alignment, annotation, model architecture and training, as well as evaluation methodology. Furthermore, we offer an in-depth discussion on two critical yet overlooked issues: The first challenges the prevalent fusion paradigm, which assumes that multi-modal fusion is applied routinely in all circumstances. Instead, we advocate for selective, rather than indiscriminate, multi-modal information, depending on the quality of the data furnished by the respective modalities. This should avoid performance degradation caused by poor quality information. The second highlights the extreme long-tail distribution in existing MMVOT datasets, which hampers generalisation. This is exemplified by the notable scarcity of animal data, limiting the advances that could be made specially in multi-modal animal tracking. Finally, we identify and outline various open issues, providing valuable suggestions for future research.

In conclusion, this survey overviews the key elements of multi-modal visual object tracking and presents a deep analysis of all aspects of multimodality in MMVOT. We hope it will serve as a valuable resource for researchers and practitioners, whether they are already familiar with MMVOT tasks or new to the field. To support further advancement of the MMVOT community as well as its deployment in smart cities, the collated statistics, supplementary material, and the details of the reported analysis are publicly available at <https://github.com/Zhangyong-Tang/MultiModal-Visual-Object-tracking> and will be continuously updated.

REFERENCES

- [1] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi *et al.*, “Agent ai: Surveying the horizons of multimodal interaction,” *arXiv preprint arXiv:2401.03568*, 2024.
- [2] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, “Lrnet: A novel representation learning guided fusion network for infrared and visible images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11 040–11 052, 2023.
- [3] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, “Object fusion tracking based on visible and infrared images: A comprehensive review,” *Inf. Fusion*, vol. 63, pp. 166–187, 2020.
- [4] Z. Tang, T. Xu, and X.-J. Wu, “A survey for deep rgbt tracking,” *arXiv preprint arXiv:2201.09296*, 2022.
- [5] Z. Zhang, J. Wang, S. Li, L. Jin, H. Wu, J. Zhao, and B. Zhang, “Review and analysis of rgbt single object tracking methods: A fusion perspective,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 8, pp. 1–27, 2024.
- [6] H. Zhang, D. Yuan, X. Shu, Z. Li, Q. Liu, X. Chang, Z. He, and G. Shi, “A comprehensive review of rgbt tracking,” *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–23, 2024.
- [7] M. Feng and J. Su, “Rgbt tracking: A comprehensive review,” *Inf. Fusion*, p. 102492, 2024.
- [8] J. Yang, Z. Li, S. Yan, F. Zheng, A. Leonardis, J.-K. Kämäräinen, and L. Shao, “Rgbt object tracking: An in-depth review,” *arXiv preprint arXiv:2203.14134*, 2022.
- [9] Z. Ou, G. Ying, D. Zhang, and Z. Zheng, “A survey of rgb-depth object tracking,” *Journal of Computer-Aided Design & Computer Graphics*, vol. 36, no. 11, pp. 1673–1690, 2024.
- [10] P. Zhang, D. Wang, and H. Lu, “Multi-modal visual tracking: Review and experimental comparison,” *Comput. Vis. Media*, vol. 10, no. 2, pp. 193–214, 2024.
- [11] C. Zhang, L. Liu, H. Wen, X. Zhou, and Y. Wang, “Awesome multi-modal object tracking,” *arXiv preprint arXiv:2405.14200*, 2024.
- [12] X. Li, J. Ren, and Y. Li, “Multi-mode filter target tracking method for mobile robot using multi-agent reinforcement learning,” *Eng. Appl. Artif. Intell.*, vol. 127, p. 107398, 2024.
- [13] C. Wu, X.-J. Wu, T. Xu, Z. Shen, and J. Kittler, “Motion complement and temporal multifocusing for skeleton-based action recognition,” *IEEE transactions on circuits and systems for video technology*, vol. 34, no. 1, pp. 34–45, 2023.
- [14] Z. Chen, T. Xu, X.-J. Wu, R. Wang, and J. Kittler, “Hybrid riemannian graph-embedding metric learning for image set classification,” *IEEE transactions on big data*, vol. 9, no. 1, pp. 75–92, 2021.
- [15] M. Kristan, J. Matas, M. Danelljan, M. Felsberg, H. J. Chang, L. Č. Zají, A. Lukežič, O. Drbohlav, Z. Zhang, K.-T. Tran *et al.*, “The first visual object tracking segmentation vots2023 challenge results,” in *Eur. Conf. Comput. Vis. Workshops*, 2023, pp. 1796–1818.
- [16] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional siamese networks for object tracking,” in *Int. Conf. Comput. Vis. Workshops*, 2016, pp. 850–865.
- [17] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, “Joint feature learning and relation modeling for tracking: A one-stream framework,” in *Eur. Conf. Comput. Vis.*, 2022, pp. 341–357.
- [18] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, “Multiple object tracking: A literature review,” *Artif. Intell.*, vol. 293, p. 103448, 2021.
- [19] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey,” *Neurocomputing*, vol. 300, pp. 17–33, 2018.
- [20] Z. Fang, S. Zhou, Y. Cui, and S. Scherer, “3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud,” *IEEE Sens. J.*, vol. 21, no. 4, pp. 4995–5011, 2020.
- [21] X. Wang, X. Shu, Z. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, “Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 13 763–13 773.
- [22] L. Peng, J. Gao, X. Liu, W. Li, S. Dong, Z. Zhang, H. Fan, and L. Zhang, “Vasttrack: Vast category visual object tracking,” in *Adv. Neural Inform. Process. Syst.*, 2024.
- [23] J. Yang, S. Gao, Z. Li, F. Zheng, and A. Leonardis, “Resource-efficient rgbt aerial tracking,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 374–13 383.
- [24] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, “Visevent: Reliable object tracking via collaboration of frame and event flows,” *IEEE T. Cybern.*, vol. 54, no. 3, pp. 1997–2010, 2023.
- [25] J. Zhang, X. Yang, Y. Fu, X. Wei, B. Yin, and B. Dong, “Object tracking by jointly exploiting frame and event domain,” in *Int. Conf. Comput. Vis.*, 2021, pp. 13 043–13 052.
- [26] Y. Zhu, X. Wang, C. Li, B. Jiang, L. Zhu, Z. Huang, Y. Tian, and J. Tang, “Crstot: Cross-resolution object tracking using unaligned frame and event cameras,” *arXiv preprint arXiv:2401.02826*, 2024.
- [27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [28] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, “Lasot: A high-quality benchmark for large-scale single object tracking,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 5374–5383.
- [29] S. Hu, D. Zhang, X. Feng, X. Li, X. Zhao, K. Huang *et al.*, “A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship,” *Adv. Neural Inform. Process. Syst.*, vol. 36, 2024.

- [30] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, "Lasher: A large-scale high-diversity benchmark for rgbt tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 392–404, 2021.
- [31] Y. Li, B. Wang, J. Sun, X. Wu, and Y. Li, "Rgb-sonar tracking benchmark and spatial cross-attention transformer tracker," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 35, no. 3, pp. 2260–2275, 2025.
- [32] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal uav tracking: A large-scale benchmark and new baseline," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8886–8895.
- [33] C. Li, T. Zhu, L. Liu, X. Si, Z. Fan, and S. Zhai, "Cross-modal object tracking: Modality-aware representations and a unified benchmark," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 1289–1296.
- [34] L. Liu, M. Zhang, C. Li, C. Li, and J. Tang, "Cross-modal object tracking via modality-aware fusion network and a large-scale dataset," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 6981–6994, 2025.
- [35] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, 2016.
- [36] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for rgb-t object tracking," in *ACM Int. Conf. Multimedia*, 2017, pp. 1856–1864.
- [37] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, p. 106977, 2019.
- [38] Z. Tang, T. Xu, Z. Feng, X. Zhu, H. Wang, P. Shao, C. Cheng, X.-J. Wu, M. Awais, S. Atito *et al.*, "Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and method," *arXiv preprint arXiv:2405.00168*, 2024.
- [39] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vision Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.
- [40] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vision Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.
- [41] X.-F. Zhu, T. Xu, Z. Tang, Z. Wu, H. Liu, X. Yang, X.-J. Wu, and J. Kittler, "Rgbdlk: A large-scale dataset and benchmark for rgb-d object tracking," in *AAAI Conf. Artif. Intell.*, vol. 37, no. 3, 2023, pp. 3870–3878.
- [42] H. Zhao, J. Chen, L. Wang, and H. Lu, "Arkittrack: a new diverse dataset for tracking using mobile rgb-d data," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 5126–5135.
- [43] S. Song and J. Xiao, "Tracking revisited using rgbd camera: Unified benchmark and baselines," in *Int. Conf. Comput. Vis.*, 2013, pp. 233–240.
- [44] S. Yan, J. Yang, J. Käpylä, F. Zheng, A. Leonardis, and J.-K. Kämäräinen, "Depthtrack: Unveiling the power of rgbd tracking," in *Int. Conf. Comput. Vis.*, 2021, pp. 10 725–10 733.
- [45] A. Lukezic, U. Kart, J. Käpylä, A. Durmush, J.-K. Kamarainen, J. Matas, and M. Kristan, "Cdtb: A color and depth visual object tracking dataset and benchmark," in *Int. Conf. Comput. Vis.*, 2019, pp. 10 013–10 022.
- [46] J. Xiao, R. Stolkin, Y. Gao, and A. Leonardis, "Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints," *IEEE T. Cybern.*, vol. 48, no. 8, pp. 2485–2499, 2017.
- [47] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 db 15 μ s latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.
- [48] Y. Wang, X. Wei, H. Shen, L. Ding, and J. Wan, "Robust fusion for rgbd tracking using cnn features," *Appl. Soft Comput.*, vol. 92, p. 106302, 2020.
- [49] Y. Wang, X. Wei, L. Luo, W. Wen, and Y. Wang, "Robust rgb-d tracking via compact cnn features," *Eng. Appl. Artif. Intell.*, vol. 96, p. 103974, 2020.
- [50] X. Zhang, X. Zhang, X. Du, X. Zhou, and J. Yin, "Learning multi-domain convolutional network for rgb-t visual tracking," in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2018, pp. 1–6.
- [51] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, "Siamft: An rgb-infrared fusion tracking method via fully convolutional siamese networks," *IEEE Access*, vol. 7, pp. 122 122–122 133, 2019.
- [52] Z. Wang, Y. Ding, P. Wu, and J. Zhang, "Wf_dimp: weight-aware dual-modal feature aggregation mechanism for rgb-t tracking," in *Symposium on Novel Photoelectronic Detection Technology and Applications*, vol. 11763, 2021, pp. 1635–1640.
- [53] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. Van De Weijer, and F. Shahbaz Khan, "Multi-modal fusion for end-to-end rgb-t tracking," in *Eur. Conf. Comput. Vis. Workshops*, 2019, pp. 0–0.
- [54] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for rgb-t object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.
- [55] Y. Wang, X. Wei, X. Tang, H. Shen, and H. Zhang, "Adaptive fusion cnn features for rgbt object tracking," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 7, pp. 7831–7840, 2021.
- [56] M. Jiang, Z. Pan, and Z. Tang, "Visual object tracking based on cross-modality gaussian-bernoulli deep boltzmann machines with rgb-d sensors," *Sensors*, vol. 17, no. 1, p. 121, 2017.
- [57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.
- [58] S. Zhai, P. Shao, X. Liang, and X. Wang, "Fast rgb-t tracking via cross-modal correlation filters," *Neurocomputing*, vol. 334, pp. 172–181, 2019.
- [59] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4277–4286.
- [60] A. Lu, Y. Guo, W. Wang, C. Li, J. Tang, and B. Luo, "Breaking shallow limits: Task-driven pixel fusion for gap-free rgbt tracking," *arXiv preprint arXiv:2503.11247*, 2025.
- [61] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, and C. N. Canagarajah, "The effect of pixel-level fusion on object tracking in multi-sensor surveillance video," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–7.
- [62] H. Zheng, N. Yuan, H. Ding, P. Hu, and Z. Yang, "Thermal infrared and visible sequences tracking via dual adversarial pixel fusion," *Multimedia Tools Appl.*, pp. 1–20, 2023.
- [63] Z. Tu, W. Pan, Y. Duan, J. Tang, and C. Li, "Rgbt tracking via reliable feature configuration," *Sci. China-Inf. Sci.*, vol. 65, no. 4, p. 142101, 2022.
- [64] L. Liu, C. Li, Y. Xiao, and J. Tang, "Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance," in *ACM Int. Conf. Multimedia*, 2023, pp. 3129–3137.
- [65] M. Zhou, X. Zhao, F. Luo, J. Luo, H. Pu, and T. Xiang, "Robust rgb-t tracking via adaptive modality weight correlation filters and cross-modality learning," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 4, pp. 1–20, 2023.
- [66] L. Feng, K. Song, J. Wang, and Y. Yan, "Exploring the potential of siamese network for rgbt object tracking," *J. Visual Commun. Image Represent.*, vol. 95, p. 103882, 2023.
- [67] F. Wang, W. Wang, L. Liu, C. Li, and J. Tang, "Siamese transformer rgbt tracking," *Appl. Intell.*, vol. 53, no. 21, pp. 24 709–24 723, 2023.
- [68] X. Lan, M. Ye, S. Zhang, and P. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
- [69] Y. Cai, X. Sui, G. Gu, and Q. Chen, "Learning modality feature fusion via transformer for rgbt-tracking," *Infrared Phys. Technol.*, vol. 133, p. 104819, 2023.
- [70] C. Guo, D. Yang, C. Li, and P. Song, "Dual siamese network for rgbt tracking via fusing predicted position maps," *The Vis. Comput.*, vol. 38, no. 7, pp. 2555–2567, 2022.
- [71] Y. Zhao, H. Lai, and G. Gao, "Rmfnet: Redetection multimodal fusion network for rgbt tracking," *Applied Sciences*, vol. 13, no. 9, p. 5793, 2023.
- [72] L. Geng, D. Zhou, K. Wang, Y. Liu, and K. Yan, "Siammgt: robust rgbt tracking via graph attention and reliable modality weight learning," *J. Supercomput.*, pp. 1–23, 2024.
- [73] C. Ó. Conaire, N. E. O'Connor, and A. Smeaton, "Thermo-visual feature fusion for object tracking using multiple spatiogram trackers," *Mach. Vision Appl.*, vol. 19, no. 5, pp. 483–494, 2008.
- [74] S. M. Ghazali and Y. Baleghi, "Rgbt tracking based on prior least absolute shrinkage and selection operator and quality aware fusion of deep and handcrafted features," *Knowledge-Based Syst.*, vol. 275, p. 110683, 2023.
- [75] X. Zhu, J. Liu, X. Xiong, and Z. Luo, "Maximize peak-to-sidelobe ratio for real-time rgbt tracking," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–4, 2024.
- [76] Z. Tang, T. Xu, H. Li, X.-J. Wu, X. Zhu, and J. Kittler, "Exploring fusion strategies for accurate rgbt visual object tracking," *Inf. Fusion*, vol. 99, p. 101881, 2023.

- [77] Y. Wang, C. Li, and J. Tang, "Learning soft-consistent correlation filters for rgb-t object tracking," in *Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018*, 2018, pp. 295–306.
- [78] Y. Li, X. Wang, Z. L. Jiang, S. Qi, X. Liu, and Q. Chen, "Rgb-d tracker under hierarchical structure," in *Conference on Computational Intelligence for Financial Engineering & Economics*, 2019, pp. 1–6.
- [79] M.-x. Jiang, C. Deng, J.-s. Shan, Y.-y. Wang, Y.-j. Jia, and X. Sun, "Hierarchical multi-modal fusion fcnn with attention model for rgb-d tracking," *Inf. Fusion*, vol. 50, pp. 1–8, 2019.
- [80] Y. Zhao, H. Lai, and G. Gao, "Hatfnet: Hierarchical adaptive trident fusion network for rgbt tracking," *Appl. Intell.*, vol. 53, no. 20, pp. 24 187–24 201, 2023.
- [81] X. Zhang, P. Ye, S. Peng, J. Liu, and G. Xiao, "Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion," *Signal Process. Image Commun.*, vol. 84, p. 115756, 2020.
- [82] Q. Xu, Y. Mei, J. Liu, and C. Li, "Multimodal cross-layer bilinear pooling for rgbt tracking," *IEEE Trans. Multimedia*, vol. 24, pp. 567–580, 2021.
- [83] W. Liu, W. Liu, and Y. Sun, "Visible-infrared dual-sensor fusion for single-object tracking," *IEEE Sens. J.*, vol. 23, no. 4, pp. 4118–4128, 2023.
- [84] A. Lu, W. Wang, C. Li, J. Tang, and B. Luo, "Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba," in *AAAI Conf. Artif. Intell.*, vol. 39, no. 6, 2025, pp. 5793–5801.
- [85] T. Luan, H. Zhang, J. Li, J. Zhang, and L. Zhuo, "Object fusion tracking for rgb-t images via channel swapping and modal mutual attention," *IEEE Sens. J.*, 2023.
- [86] Y. Zhu, C. Li, X. Wang, J. Tang, and Z. Huang, "Rgbt tracking via progressive fusion transformer with dynamically guided learning," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 9, pp. 8722–8735, 2024.
- [87] B. Xu, Y. Xu, R. Hou, J. Bei, T. Ren, and G. Wu, "Rgb-d tracking via hierarchical modality aggregation and distribution network," in *ACM Int. Conf. Multimedia Asia*, 2023, pp. 1–7.
- [88] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, "Deep adaptive fusion network for high performance rgbt tracking," in *Eur. Conf. Comput. Vis. Workshops*, 2019, pp. 0–0.
- [89] Y. Li, H. Lai, L. Wang, and Z. Jia, "Multibranch adaptive fusion network for rgbt tracking," *IEEE Sens. J.*, vol. 22, no. 7, pp. 7084–7093, 2022.
- [90] Z. Yang, P. Huang, D. He, Z. Cai, and Z. Yin, "Siammmf: multi-modal multi-level fusion object tracking based on siamese networks," *Mach. Vision Appl.*, vol. 34, no. 1, p. 7, 2023.
- [91] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust rgbt tracking," *IEEE T. Intell. Veh.*, vol. 6, no. 1, pp. 121–130, 2020.
- [92] J. Mei, D. Zhou, J. Cao, R. Nie, and Y. Guo, "Hdinet: Hierarchical dual-sensor interaction network for rgbt tracking," *IEEE Sens. J.*, vol. 21, no. 15, pp. 16 915–16 926, 2021.
- [93] C. Zhao, B. Mo, D. Li, X. Wang, J. Zhao, and J. Xu, "Mcssafnet: A multi-scale state-space attention fusion network for rgbt tracking," *Opt. Commun.*, vol. 577, p. 131394, 2025.
- [94] K. Yan, C. Wang, D. Zhou, and Z. Zhou, "Rgbt tracking via multi-stage matching guidance and context integration," *Neural Process. Lett.*, vol. 55, no. 8, pp. 11 073–11 087, 2023.
- [95] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 10, pp. 3727–3739, 2020.
- [96] Z. Zhang, B. Zeng, X. Ni, and Y. Du, "Btmtrack: Robust rgb-t tracking via dual-template bridging and temporal-modal candidate elimination," *arXiv preprint arXiv:2501.03616*, 2025.
- [97] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, and J. Liu, "Temporal adaptive rgbt tracking with modality prompt," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 6, 2024, pp. 5436–5444.
- [98] Z. Zeng, X. Liu, M. Sun, H. Wang, and J. Liu, "Cross fusion rgb-t tracking with bi-directional adapter," *arXiv preprint arXiv:2408.16979*, 2024.
- [99] G. Ying, D. Zhang, Z. Ou, X. Wang, and Z. Zheng, "Temporal adaptive bidirectional bridging for rgb-d tracking," *Pattern Recognit.*, p. 111053, 2024.
- [100] D. Sun, Y. Pan, A. Lu, C. Li, and B. Luo, "Transformer rgbt tracking with spatio-temporal multimodal tokens," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 11, pp. 12 059–12 072, 2024.
- [101] H. Shi, X. Mu, H. He, C. Zhong, B. Zhang, and P. Zhao, "Iamtrack: interframe appearance and modality tokens propagation with temporal modeling for rgbt tracking," *Appl. Intell.*, vol. 55, no. 7, p. 583, 2025.
- [102] H. Xing, W. Wei, L. Zhang, and Y. Zhang, "Multi-scale feature extraction and fusion with attention interaction for rgb-t tracking," *Pattern Recognit.*, p. 110917, 2024.
- [103] T. Zhang, Q. Jiao, Q. Zhang, and J. Han, "Exploring multi-modal spatial-temporal contexts for high-performance rgb-t tracking," *IEEE Trans. Image Process.*, vol. 33, pp. 4303–4318, 2024.
- [104] F. Zhang, H. Peng, L. Yu, Y. Zhao, and B. Chen, "Dual-modality space-time memory network for rgbt tracking," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.
- [105] X. Liu and Z. Lei, "Dynamic tracking aggregation with transformers for rgb-t tracking," *J. Inf. Process. Syst.*, vol. 19, no. 1, pp. 80–88, 2023.
- [106] L. Liu, C. Li, F. Wang, L. Shen, and J. Tang, "Prototype-based cross-modal object tracking," *Inf. Fusion*, vol. 118, p. 102941, 2025.
- [107] C. Bian and S.-I. Kamata, "Mask refined deep fusion network with dynamic memory for robust rgbt tracking," in *IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence*, 2023, pp. 117–124.
- [108] Z. Tang, T. Xu, X.-J. Wu, and J. Kittler, "Temporal aggregation for real-time rgbt tracking via fast decision-level fusion," *Pattern Recognit. Lett.*, vol. 193, pp. 29–35, 2025.
- [109] M. Feng, K. Song, Y. Wang, J. Liu, and Y. Yan, "Learning discriminative update adaptive spatial-temporal regularized correlation filter for rgb-t tracking," *J. Visual Commun. Image Represent.*, vol. 72, p. 102881, 2020.
- [110] C. Wang, C. Xu, Z. Cui, L. Zhou, T. Zhang, X. Zhang, and J. Yang, "Cross-modal pattern-propagation for rgb-t tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 7064–7073.
- [111] S. Lai, C. Liu, J. Zhu, B. Kang, Y. Liu, D. Wang, and H. Lu, "Mambavt: Spatio-temporal contextual modeling for robust rgb-t tracking," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.
- [112] L. Jun, L. Zhongqiang, and X. Xingzhong, "Rgb-t long-term tracking algorithm via local sampling and global proposals," *Signal Image Video Process.*, vol. 16, no. 8, pp. 2221–2229, 2022.
- [113] J. Liu, Z. Luo, and X. Xiong, "Online learning samples and adaptive recovery for robust rgb-t tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 2, pp. 724–737, 2023.
- [114] K. Meshgi, S.-i. Maeda, S. Oba, H. Skibbe, Y.-z. Li, and S. Ishii, "An occlusion-aware particle filter tracker to handle complex and persistent occlusions," *Comput. Vision Image Understanding*, vol. 150, pp. 81–94, 2016.
- [115] N. Park and S. Kim, "How do vision transformers work?" *arXiv preprint arXiv:2202.06709*, 2022.
- [116] J. Mei, J. Zhou, J. Wang, J. Hao, D. Zhou, and J. Cao, "Learning multifrequency integration network for rgbt tracking," *IEEE Sens. J.*, vol. 24, no. 9, pp. 15 517–15 530, 2024.
- [117] Z. Gao, D. Zhou, J. Cao, Y. Liu, and Q. Shan, "Fdtrack: A dual-head focus tracking network with frequency enhancement," *IEEE Sens. J.*, vol. 25, no. 2, pp. 3879–3897, 2025.
- [118] L. Lei and X. Li, "Rgb-t tracking with frequency hybrid awareness," *Image Vis. Comput.*, vol. 152, p. 105330, 2024.
- [119] S. Chen, L. Xu, X. Li, and C. Tian, "Frequency-space enhanced and temporal adaptative rgbt object tracking," *Neurocomputing*, p. 130240, 2025.
- [120] J. Zhang, J. Yang, Z. Liu, and J. Wang, "Rgbt tracking via frequency-aware feature enhancement and unidirectional mixed attention," *Neurocomputing*, p. 128908, 2024.
- [121] Y. Li, B. Wang, Y. Li, Z. Yu, and L. Wang, "Transformer-based rgbt tracking with channel and spatial feature fusion," *arXiv preprint arXiv:2405.03177*, 2024.
- [122] S. Zhai, Y. Wu, L. Liu, and J. Tang, "Rgbt tracking based on modality feature enhancement," *Multimedia Tools Appl.*, vol. 83, no. 10, pp. 29 311–29 330, 2024.
- [123] R. Hou, B. Xu, T. Ren, and G. Wu, "Mtnet: learning modality-aware representation with transformer for rgbt tracking," in *nt. Conf. Multimedia and Expo*, 2023, pp. 1163–1168.
- [124] R. Shi, C. Wang, G. Zhao, and C. Xu, "Sca-mma: Spatial and channel-aware multi-modal adaptation for robust rgb-t object tracking," *Electronics*, vol. 11, no. 12, p. 1820, 2022.
- [125] L. Chen, B. Zhong, Q. Liang, Y. Zheng, Z. Mo, and S. Song, "Top-down cross-modal guidance for robust rgb-t tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 12, pp. 12 388–12 398, 2024.

- [126] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang, "Learning target-oriented dual attention for robust rgb-t tracking," *arXiv preprint arXiv:1908.04441*, 2019.
- [127] M. Feng and J. Su, "Learning reliable modal weight with transformer for robust rgbt tracking," *Knowledge-Based Syst.*, vol. 249, p. 108945, 2022.
- [128] C. Xu, Z. Cui, C. Wang, C. Zhou, and J. Yang, "Learning cross-modal interaction for rgb-t tracking," *Sci. China-Inf. Sci.*, vol. 66, no. 1, p. 119103, 2023.
- [129] B. Kang, D. Liang, J. Mei, X. Tan, Q. Zhou, and D. Zhang, "Robust rgb-t tracking via graph attention-based bilinear pooling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9900–9911, 2022.
- [130] R. Yang, X. Wang, Y. Zhu, and J. Tang, "Deep triply attention network for rgbt tracking," *Cognit. Comput.*, vol. 15, no. 6, pp. 1934–1946, 2023.
- [131] Z. Gao, D. Li, G. Wen, Y. Kuai, and R. Chen, "Drone based rgbt tracking with dual-feature aggregation network," *Drones*, vol. 7, no. 9, p. 585, 2023.
- [132] R. Yang, X. Wang, C. Li, J. Hu, and J. Tang, "Rgbt tracking via cross-modality message passing," *Neurocomputing*, vol. 462, pp. 365–375, 2021.
- [133] X. Liu, Y. Luo, K. Yan, J. Chen, and Z. Lei, "Cmc2r: Cross-modal collaborative contextual representation for rgbt tracking," *IET Image Proc.*, vol. 16, no. 5, pp. 1500–1510, 2022.
- [134] R. Hou, T. Ren, and G. Wu, "Mimet: A robust rgbt tracking jointly with multi-modal interaction and refinement," in *nt. Conf. Multimedia and Expo*, 2022, pp. 1–6.
- [135] T. Zhang, X. He, Y. Luo, Q. Zhang, and J. Han, "Exploring target-related information with reliable global pixel relationships for robust rgb-t tracking," *Pattern Recognit.*, vol. 155, p. 110707, 2024.
- [136] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, and S. Liu, "Bridging search region interaction with template for rgb-t tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 630–13 639.
- [137] Y. Xiao, J. Zhao, A. Lu, C. Li, B. Yin, Y. Lin, and C. Liu, "Cross-modulated attention transformer for rgbt tracking," in *AAAI Conf. Artif. Intell.*, vol. 39, no. 8, 2025, pp. 8682–8690.
- [138] C. Li, S. Hu, S. Gao, and J. Tang, "Real-time grayscale-thermal tracking via laplacian sparse representation," in *MultiMedia Modeling: 22nd International Conference*, 2016, pp. 54–65.
- [139] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Int. Conf. Inf. Fusion*, 2011, pp. 1–8.
- [140] C. Li, X. Sun, X. Wang, L. Zhang, and J. Tang, "Grayscale-thermal object tracking via multitask laplacian sparse representation," *IEEE Trans. Syst. Man Cybern. -Syst.*, vol. 47, no. 4, pp. 673–681, 2017.
- [141] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Sci. China-Inf. Sci.*, vol. 55, pp. 590–599, 2012.
- [142] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Int. Conf. Inf. Fusion*, 2011, pp. 1–8.
- [143] C. Li, C. Zhu, S. Zheng, B. Luo, and J. Tang, "Two-stage modality-graphs regularized manifold ranking for rgb-t tracking," *Signal Process. Image Commun.*, vol. 68, pp. 207–217, 2018.
- [144] C. Li, Z. Xiang, J. Tang, B. Luo, and F. Wang, "Rgbt tracking via noise-robust cross-modal ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 5019–5031, 2021.
- [145] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking," in *Eur. Conf. Comput. Vis.*, 2018, pp. 808–823.
- [146] B. Kang, L. Liu, S. Zhao, and S. Du, "Robust rgb-t tracking via consistency regulated scene perception," in *IEEE Int. Conf. Image Process.*, 2023, pp. 510–514.
- [147] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, "Learning local-global multi-graph descriptors for rgb-t object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 29, no. 10, pp. 2913–2926, 2018.
- [148] P. Chen, S. Gong, W. Ying, X. Du, and S. Zhong, "Tefnet: Target-aware enhanced fusion network for rgb-t tracking," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 432–443.
- [149] Z. Yu, H. Fan, Q. Wang, Z. Li, and Y. Tang, "Region selective fusion network for robust rgbt tracking," *IEEE Signal Process. Lett.*, vol. 30, pp. 1357–1361, 2023.
- [150] Y. Kuai, G. Wen, D. Li, and J. Xiao, "Target-aware correlation filter tracking in rgbd videos," *IEEE Sens. J.*, vol. 19, no. 20, pp. 9522–9531, 2019.
- [151] Y. Chen and L. Wang, "emoe-tracker: Environmental moe-based transformer for robust event-guided object tracking," *IEEE Rob. Autom. Lett.*, vol. 10, no. 2, pp. 1393–1400, 2025.
- [152] Z. Zhu, J. Hou, and D. O. Wu, "Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers," in *Int. Conf. Comput. Vis.*, 2023, pp. 22 045–22 055.
- [153] Y. Cai, X. Sui, G. Gu, and Q. Chen, "Multi-modal interaction with token division strategy for rgb-t tracking," *Pattern Recognit.*, p. 110626, 2024.
- [154] Q. Zhang, X. Liu, and T. Zhang, "Rgb-t tracking by modality difference reduction and feature re-selection," *Image Vis. Comput.*, vol. 127, p. 104547, 2022.
- [155] L. Zhao, M. Zhu, H. Ren, and L. Xue, "Channel exchanging for rgb-t tracking," *Sensors*, vol. 21, no. 17, p. 5800, 2021.
- [156] Y. Luo, X. Guo, H. Feng, and L. Ao, "Rgb-t tracking via multi-modal mutual prompt learning," *arXiv preprint arXiv:2308.16386*, 2023.
- [157] B. Cao, J. Guo, P. Zhu, and Q. Hu, "Bi-directional adapter for multimodal tracking," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 2, 2024, pp. 927–935.
- [158] Z. Ou, D. Zhang, G. Ying, and Z. Zheng, "Ubpt: Unidirectional and bidirectional prompts for rgbd tracking," *IEEE Sens. J.*, vol. 24, no. 22, pp. 37 503–37 513, 2024.
- [159] H. Fan, Z. Yu, Q. Wang, B. Fan, and Y. Tang, "Querytrack: Joint-modality query fusion network for rgbt tracking," *IEEE Trans. Image Process.*, vol. 33, pp. 3187–3199, 2024.
- [160] Y. Du, B. Zeng, Q. Wei, B. Zhang, and H. Hu, "Transformer-mamba-based trident-branch rgb-t tracker," in *Pacific Rim International Conference on Artificial Intelligence*, 2024, pp. 27–40.
- [161] J. Zhang, Y. Qin, S. Fan, Z. Xiao, and J. Zhang, "Siamtfa: Siamese triple-stream feature aggregation network for efficient rgbt tracking," *IEEE Trans. Intell. Transport. Syst.*, vol. 26, no. 2, pp. 1900–1913, 2025.
- [162] Y. Liu, D. Zhou, J. Cao, K. Yan, and L. Geng, "Specific and collaborative representations siamese network for rgbt tracking," *IEEE Sens. J.*, vol. 24, no. 11, pp. 18 520–18 534, 2024.
- [163] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, "Rgbt tracking by trident fusion network," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 32, no. 2, pp. 579–592, 2021.
- [164] F. Zhang, S. Ma, Z. Li, and Y. Zhang, "Rgb-t tracking via multi-expert correlation filters using spatial-temporal robustness," in *International Conference on Mechanical, Control and Computer Engineering*, 2020, pp. 360–364.
- [165] S. Gao, J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Learning dual-fused modality-aware representations for rgbd tracking," in *Eur. Conf. Comput. Vis.*, 2022, pp. 478–494.
- [166] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for rgbt tracking," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2831–2838.
- [167] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware rgbt tracking," in *Eur. Conf. Comput. Vis.*, 2020, pp. 222–237.
- [168] L. Liu, C. Li, Y. Xiao, R. Ruan, and M. Fan, "Rgbt tracking via challenge-based appearance disentanglement and interaction," *IEEE Trans. Image Process.*, vol. 33, pp. 1753–1767, 2024.
- [169] C. Li, T. Wang, Z. Ding, Y. Xiao, and J. Tang, "Dynamic disentangled fusion network for rgbt tracking," *arXiv preprint arXiv:2412.08441*, 2024.
- [170] W. Wang, C. Li, D. Zhang, H. Zhou, M. Xie, H. Zhou, and K. Fu, "Fcfnet: A challenge-based feature complementary fusion network for rgbt tracking," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 2239–2251, 2025.
- [171] C. Long Li, A. Lu, A. Hua Zheng, Z. Tu, and J. Tang, "Multi-adapter rgbt tracking," in *Eur. Conf. Comput. Vis. Workshops*, 2019, pp. 0–0.
- [172] Y. Luo, X. Guo, M. Dong, and J. Yu, "Learning modality complementary features with mixed attention mechanism for rgb-t tracking," *Sensors*, vol. 23, no. 14, p. 6609, 2023.
- [173] L. Chen, Y. Huang, H. Li, Z. Zhou, and Z. He, "Simplifying cross-modal interaction via modality-shared features for rgbt tracking," in *ACM Int. Conf. Multimedia*, 2024, pp. 1573–1582.
- [174] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "Rgbt tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5613–5625, 2021.
- [175] Z. Tu, C. Lin, W. Zhao, C. Li, and J. Tang, "M 5 l: multi-modal multi-margin metric learning for rgbt tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 85–98, 2021.
- [176] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in rgb-t videos using modal-aware attention network and competitive learning," *Sensors*, vol. 20, no. 2, p. 393, 2020.

- [177] A. Lu, J. Zhao, C. Li, Y. Xiao, and B. Luo, "Breaking modality gap in rgbt tracking: Coupled knowledge distillation," in *ACM Int. Conf. Multimedia*, 2024, pp. 9291–9300.
- [178] Y. Chen, M. Wan, Y. Xu, X. Zhang, Q. Chen, and G. Gu, "Residual learning-based two-stream network for rgb-t object tracking," *J. Electron. Imaging*, vol. 31, no. 6, pp. 063 041–063 041, 2022.
- [179] X. Liu, Y. Luo, Y. Zhang, and Z. Lei, "Semantic-guided fusion for multiple object tracking and rgb-t tracking," *IET Image Proc.*, vol. 17, no. 11, pp. 3281–3291, 2023.
- [180] T. Zhang, K. Debattista, Q. Zhang, J. Han *et al.*, "Revisiting motion information for rgb-event tracking with mot philosophy," vol. 37, 2024, pp. 89 346–89 372.
- [181] S. Fan, C. He, C. Wei, Y. Zheng, and X. Chen, "Bayesian dumbbell diffusion model for rgbt object tracking with enriched priors," *IEEE Signal Process. Lett.*, vol. 30, pp. 873–877, 2023.
- [182] F. Li, W. Xia, D. Zhou, and J. Cao, "Mkfracker: An rgbt tracker via multimodal knowledge embedding and feature interaction," *Knowledge-Based Syst.*, vol. 310, p. 112860, 2025.
- [183] J. Wang, F. Liu, L. Jiao, Y. Gao, H. Wang, S. Li, L. Li, P. Chen, and X. Liu, "Visual and language collaborative learning for rgbt object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 12, pp. 12 770–12 781, 2024.
- [184] Y. Cai, X. Sui, and G. Gu, "Multi-modal multi-task feature fusion for rgbt tracking," *Inf. Fusion*, vol. 97, p. 101816, 2023.
- [185] C. Guo and L. Xiao, "High speed and robust rgb-thermal tracking via dual attentive stream siamese network," in *IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 803–806.
- [186] Y. Xue, J. Zhang, Z. Lin, C. Li, B. Huo, and Y. Zhang, "Siamcaf: Complementary attention fusion-based siamese network for rgbt tracking," *Remote Sensing*, vol. 15, no. 13, p. 3252, 2023.
- [187] K. Qi, L. Chen, Y. Zhou, and Y. Qi, "Multi-modal fusion object tracking based on fully convolutional siamese network," in *Asia Conference on Algorithms, Computing and Machine Learning*, 2023, pp. 440–444.
- [188] J. Peng, H. Zhao, Z. Hu, Y. Zhuang, and B. Wang, "Siamese infrared and visible light fusion network for rgb-t tracking," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 9, pp. 3281–3293, 2023.
- [189] G. Wang, Q. Jiang, X. Jin, Y. Lin, Y. Wang, and W. Zhou, "Siamtdr: Time-efficient rgbt tracking via disentangled representations," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 1, pp. 167–181, 2023.
- [190] X. Li, J. Cai, W. Ding, and Z. Fang, "Robust rgb-t tracking via bimodal constrained correlation filtering," in *IEEE International Conference on Progress in Informatics and Computing*, 2020, pp. 159–162.
- [191] Y. Huang, X. Li, R. Lu, and N. Qi, "Rgb-t object tracking via sparse response-consistency discriminative correlation filters," *Infrared Phys. Technol.*, vol. 128, p. 104509, 2023.
- [192] Q. Zhao, J. Liu, J. Wang, and X. Xiong, "Real-time rgbt target tracking based on attention mechanism," *Electronics*, vol. 13, no. 13, p. 2517, 2024.
- [193] Y. Luo, X. Guo, and H. Li, "From two stream to one stream: Efficient rgb-t tracking via mutual prompt learning and knowledge distillation," *arXiv preprint arXiv:2403.16834*, 2024.
- [194] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient rgbt tracking via cross-modality distillation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 5404–5413.
- [195] J. Xia, D. Shi, K. Song, L. Song, X. Wang, S. Jin, L. Zhou, Y. Cheng, L. Jin, Z. Zhu *et al.*, "Unified single-stage transformer network for efficient rgb-t tracking," *arXiv preprint arXiv:2308.13764*, 2023.
- [196] J. Huang, S. Wang, S. Wang, Z. Wu, X. Wang, and B. Jiang, "Mamba-fetrack: Frame-event tracking via state space model," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2024, pp. 3–18.
- [197] T. Zhang, X. He, Q. Jiao, Q. Zhang, and J. Han, "Amnet: Learning to align multi-modality for rgb-t tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 8, pp. 7386–7400, 2024.
- [198] S. Qing, W. Yifan, G. Yu, and M. Yang, "Unveiling the power of unpaired multi-modal data for rgbt tracking," in *International Conference on Artificial Intelligence and Electromechanical Automation*, vol. 12709, 2023, pp. 672–677.
- [199] X. Hu, B. Zhong, Q. Liang, S. Zhang, N. Li, and X. Li, "Toward modalities correlation for rgb-t tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 10, pp. 9102–9111, 2024.
- [200] A. Lu, C. Li, J. Zhao, J. Tang, and B. Luo, "Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks," *Int. J. Comput. Vis.*, pp. 1–21, 2024.
- [201] S. Li, R. Yao, Y. Zhou, H. Zhu, B. Liu, J. Zhao, and Z. Shao, "Un-supervised rgb-t object tracking with attentional multi-modal feature fusion," *Multimedia Tools Appl.*, vol. 82, no. 15, pp. 23 595–23 613, 2023.
- [202] S. Li, R. Yao, Y. Zhou, H. Zhu, J. Zhao, Z. Shao, and A. El Saddik, "Motion-aware self-supervised rgbt tracking with multi-modality hierarchical transformers," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 12, pp. 1–23, 2024.
- [203] X.-F. Zhu, T. Xu, S. Atito, M. Awais, X.-J. Wu, Z. Feng, and J. Kittler, "Self-supervised learning for rgb-d object tracking," *Pattern Recognit.*, p. 110543, 2024.
- [204] R. Ramon, H. Cohen-Duwek, and E. E. Tsur, "Ed-dcfnet: An unsupervised encoder-decoder neural model for event-driven feature extraction and object tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 2191–2199.
- [205] M.-x. Jiang, C. Deng, M.-m. Zhang, J.-s. Shan, and H. Zhang, "Multimodal deep feature fusion (mmdff) for rgb-d tracking," *Complexity*, vol. 2018, no. 1, p. 5676095, 2018.
- [206] S. Fan, X. Chen, C. He, L. Yu, Z. Mao, and Y. Zheng, "Multiple frequency-spatial network for rgbt tracking in the presence of motion blur," *Neural Comput. Appl.*, vol. 35, no. 34, pp. 24 389–24 406, 2023.
- [207] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust rgb-t tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3335–3347, 2021.
- [208] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, "Duality-gated mutual condition network for rgbt tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4118–4131, 2025.
- [209] S. Pan, H. Wang, D. Li, Y. Zhang, B. Shiragapur, X. Liu, and Q. Yu, "A lightweight robust rgb-t object tracker based on jitter factor and associated kalman filter," *Inf. Fusion*, p. 102842, 2024.
- [210] Z. Gao, D. Zhou, J. Cao, Y. Liu, and Q. Shan, "Enhanced rgbt tracking network with semantic generation and historical context," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–17, 2025.
- [211] F. Gao, W. Wu, Y. Jin, J. Tang, H. Zheng, S. Ma, and J. Yu, "Tvttracker: Target-adaptive text-guided visual fusion for multi-modal rgb-t tracking," *IEEE Internet Things J.*, pp. 1–1, 2025.
- [212] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, 2009.
- [213] Z. Feng, L. Yan, J. Bai, Y. Xia, and B. Xiao, "A content-aware correlation filter with multi-feature fusion for rgb-t tracking," *J. Syst. Eng. Electron.*, vol. 35, no. 6, pp. 1357–1371, 2024.
- [214] Z. Zhao, W. Li, and F. Pan, "An rgbt object tracking method for solving camera motion based on correlation filter," in *Chinese Control and Decision Conference*, 2023, pp. 3526–3531.
- [215] C. Luo, B. Sun, K. Yang, T. Lu, and W.-C. Yeh, "Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme," *Infrared Phys. Technol.*, vol. 99, pp. 265–276, 2019.
- [216] B. Li, F. Peng, T. Hui, X. Wei, X. Wei, L. Zhang, H. Shi, and S. Liu, "Rgb-t tracking with template-bridged search interaction and target-preserved template updating," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 634–649, 2025.
- [217] J. Mei, Y. Liu, C. Wang, D. Zhou, R. Nie, and J. Cao, "Asymmetric global-local mutual integration network for rgbt tracking," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–17, 2022.
- [218] Z. Gao, D. Zhou, K. Yan, and Y. Liu, "Real-time rgbt tracking via isometric feature encoding networking," *Signal Image Video Process.*, vol. 19, no. 1, p. 16, 2025.
- [219] J. Yang, E. Dong, J. Tong, S. Yang, Z. Zhang, and W. Li, "Differential enhancement and commonality fusion for rgbt tracking," in *IEEE International Conference on Mechatronics and Automation*, 2023, pp. 351–356.
- [220] X.-F. Zhu, T. Xu, and X.-J. Wu, "Adaptive colour-depth aware attention for rgb-d object tracking," *IEEE Signal Process. Lett.*, 2024.
- [221] M. Wang, F. Shi, X. Cheng, and S. Chen, "Prior knowledge-driven hybrid prompter learning for rgb-event tracking," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.
- [222] G. Zhang, Q. Liang, Z. Mo, N. Li, and B. Zhong, "Visual adapt for rgbd tracking," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 9391–9395.
- [223] Y. Liu, Z. Gao, Y. Cao, and D. Zhou, "Two-stage unidirectional fusion network for rgbt tracking," *Knowledge-Based Syst.*, vol. 310, p. 112983, 2025.
- [224] J. Zhang, Y. Wang, W. Liu, M. Li, J. Bai, B. Yin, and X. Yang, "Frame-event alignment and fusion network for high frame rate tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9781–9790.

- [225] Z. Tang, T. Xu, X.-J. Wu, and J. Kittler, "Multi-level fusion for robust rgbt tracking via enhanced thermal representation," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 10, pp. 1–24, 2024.
- [226] H. Li, Y. Cao, L. Guo, Q. Chen, Z. Ding, and S. Xie, "Fadsiamnet: feature affinity drift siamese network for rgb-t target tracking," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 4, pp. 2779–2799, 2025.
- [227] H. Wang, T. Xu, Z. Tang, X.-J. Wu, and J. Kittler, "Multi-modal adapter for rgb-t tracking," *Inf. Fusion*, vol. 118, p. 102940, 2025.
- [228] J. Li, H. Wu, Y. Gu, J. Lu, and X. Sun, "Dusiamie: A lightweight multidimensional infrared-enhanced rgbt tracking algorithm for edge device deployment," *Electronics*, vol. 13, no. 23, p. 4721, 2024.
- [229] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time rgb-t tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 2714–2729, 2021.
- [230] X. SHAO, H. Li, Z. Lu, M. Bo, H. ZeHui *et al.*, "Special attribute-based cross-modal interactive fusion network for rgbt tracking," 2024.
- [231] X. Shao, H. Li, Z. Lu, M. Bo, H. Zehui *et al.*, "Special attribute-based cross-modal interactive fusion network for rgbt tracking," 2024.
- [232] P. Shao, T. Xu, Z. Tang, L. Li, X.-J. Wu, and J. Kittler, "Tenet: targetness entanglement incorporating with multi-scale pooling and mutually-guided fusion for rgb-e object tracking," *Neural Networks*, vol. 183, p. 106948, 2025.
- [233] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural networks*, vol. 111, pp. 47–63, 2019.
- [234] Y. Ji, Q. Zhao, Y. Liang, and J. Wu, "Snntrack: Spiking neural network based prompt for high-accuracy rgbe tracking," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025, pp. 1–5.
- [235] J. Zhang, K. Zhao, B. Dong, Y. Fu, Y. Wang, X. Yang, and B. Yin, "Multi-domain collaborative feature representation for robust visual object tracking," *The Vis. Comput.*, vol. 37, no. 9, pp. 2671–2683, 2021.
- [236] S. Liu, G. Wang, Y. Song, J. Huang, Y. Huang, Y. Zhou, and S. Wang, "Siameft: adaptive-time feature extraction hybrid network for rgbe multi-domain object tracking," *Front. Neurosci.*, vol. 18, p. 1453419, 2024.
- [237] H. Sun, R. Liu, W. Cai, J. Wang, Y. Wang, H. Tang, Y. Cui, D. Yao, and D. Guo, "Reliable object tracking by multimodal hybrid feature extraction and transformer-based fusion," *arXiv preprint arXiv:2405.17903*, 2024.
- [238] P. Shao, T. Xu, X.-F. Zhu, X.-J. Wu, and J. Kittler, "Dynamic subframe splitting and spatio-temporal motion entangled sparse attention for rgb-e tracking," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2024, pp. 107–121.
- [239] G. M. García, D. A. Klein, J. Stückler, S. Frintrop, and A. B. Cremers, "Adaptive multi-cue 3d tracking of arbitrary objects," in *Pattern Recognit.*, 2012, pp. 357–366.
- [240] J. Leng and Y. Liu, "Real-time rgb-d visual tracking with scale estimation and occlusion handling," *IEEE Access*, vol. 6, pp. 24 256–24 263, 2018.
- [241] U. Kart, J.-K. Kämäräinen, J. Matas, L. Fan, and F. Cricri, "Depth masked discriminative correlation filter," in *Int. Conf. Pattern Recog.*, 2018, pp. 2112–2117.
- [242] G. Li, L. Huang, P. Zhang, Q. Li, and Y. Huo, "Depth information aided constrained correlation filter for visual tracking," in *IOP Conference Series: Earth and Environmental Science*, vol. 234, no. 1, 2019, p. 012005.
- [243] F. Xiao, Q. Wu, and H. Huang, "Single-scale siamese network based rgb-d object tracking with adaptive bounding boxes," *Neurocomputing*, vol. 451, pp. 192–204, 2021.
- [244] H. Shi, C. Gao, and N. Sang, "Using consistency of depth gradient to improve visual tracking in rgb-d sequences," in *Chinese Automation Congress*, 2015, pp. 518–522.
- [245] Y. Liu, X.-Y. Jing, J. Nie, H. Gao, J. Liu, and G.-P. Jiang, "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 664–677, 2018.
- [246] Y. Qian, S. Yan, A. Lukežič, M. Kristan, J.-K. Kämäräinen, and J. Matas, "Dal: A deep depth-aware long-term tracker," in *Int. Conf. Pattern Recog.*, 2021, pp. 7825–7832.
- [247] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.
- [248] P. Zhao, Q. Liu, W. Wang, and Q. Guo, "Tsdm: Tracking by siamrpn++ with a depth-refiner and a mask-generator," in *Int. Conf. Pattern Recog.*, 2021, pp. 670–676.
- [249] Z.-a. Ma and Z.-y. Xiang, "Robust object tracking with rgbd-based sparse learning," *Front. Inf. Technol. Electron. Eng.*, vol. 18, no. 7, pp. 989–1001, 2017.
- [250] M. Camplani, S. L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling," in *BMVC*, vol. 3, 2015, pp. 01–12.
- [251] S. Hannuna, M. Camplani, J. Hall, M. Mirmehdi, D. Damen, T. Burghardt, A. Paiement, and L. Tao, "Ds-kcf: a real-time tracker for rgb-d data," *J. Real-Time Image Process.*, vol. 16, pp. 1439–1458, 2019.
- [252] P. Ding and Y. Song, "Robust object tracking using color and depth images with a depth based occlusion handling and recovery," in *International conference on fuzzy systems and knowledge discovery*, 2015, pp. 930–935.
- [253] Y. Zhai, P. Song, Z. Mou, X. Chen, and X. Liu, "Occlusion-aware correlation particle filter target tracking based on rgbd data," *IEEE Access*, vol. 6, pp. 50 752–50 764, 2018.
- [254] X.-F. Zhu, T. Xu, X.-J. Wu, and J. Kittler, "Feature enhancement and coarse-to-fine detection for rgb-d tracking," *Pattern Recognit. Lett.*, vol. 179, pp. 130–136, 2024.
- [255] W. Zhang, K. Yang, Y. Xin, and R. Meng, "An occlusion-aware rgbd visual object tracking method based on siamese network," in *IEEE International Conference on Signal Processing*, vol. 1, 2020, pp. 327–332.
- [256] H. Zhang, M. Cai, and J. Li, "A real-time rgb-d tracker based on kcf," in *Chinese Control And Decision Conference*, 2018, pp. 4856–4861.
- [257] A. Benou, I. Benou, and R. Hagage, "Occlusion handling method for object tracking using rgb-d data," in *IEEE 28th Convention of Electrical & Electronics Engineers in Israel*, 2014, pp. 1–5.
- [258] U. Kart, A. Lukežic, M. Kristan, J.-K. Kamarainen, and J. Matas, "Object tracking by reconstruction with view-specific discriminative correlation filters," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1339–1348.
- [259] N. An, X.-G. Zhao, and Z.-G. Hou, "Online rgb-d tracking via detection-learning-segmentation," in *Int. Conf. Pattern Recog.*, 2016, pp. 1231–1236.
- [260] B. Li, Y. Zhuge, S. Jiang, L. Wang, Y. Wang, and H. Lu, "3d prompt learning for rgb-d tracking," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 2527–2544.
- [261] A. Bibi, T. Zhang, and B. Ghanem, "3d part-based sparse tracker with automatic synchronization and registration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1439–1448.
- [262] Y. Xie, Y. Lu, and S. Gu, "Rgb-d object tracking with occlusion detection," in *International Conference on Computational Intelligence and Security*, 2019, pp. 11–15.
- [263] Y. Zheng, B. Zhong, Q. Liang, G. Li, R. Ji, and X. Li, "Toward unified token learning for vision-language tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 4, pp. 2125–2135, 2024.
- [264] X. Liu, L. Zhou, Z. Zhou, J. Chen, and Z. He, "Mambavt: Time-evolving multimodal state space model for vision-language tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 8731–8741.
- [265] Q. Feng, V. Ablavsky, Q. Bai, G. Li, and S. Sclaroff, "Real-time visual object tracking with natural language description," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 700–709.
- [266] J. Ge, X. Chen, J. Cao, X. Zhu, W. Liu, and B. Liu, "Beyond visual cues: Synchronously exploring target-centric semantics for vision-language tracking," *arXiv preprint arXiv:2311.17085*, 2023.
- [267] H. Zhu, Q. Lu, L. Xue, G. Yuan, and K. Zhang, "Joint feature extraction and alignment in object tracking with vision-language model," *Eng. Appl. Artif. Intell.*, vol. 152, p. 110787, 2025.
- [268] D. Ma and X. Wu, "Tracking by natural language specification with long short-term context decoupling," in *Int. Conf. Comput. Vis.*, 2023, pp. 14 012–14 021.
- [269] X. Li, Y. Huang, Z. He, Y. Wang, H. Lu, and M.-H. Yang, "Citetracker: Correlating image and text for visual tracking," in *Int. Conf. Comput. Vis.*, 2023, pp. 9974–9983.
- [270] P. Ye, G. Xiao, and J. Liu, "Multimodal features alignment for vision-language object tracking," *Remote Sensing*, vol. 16, no. 7, p. 1168, 2024.
- [271] Q. Feng, V. Ablavsky, Q. Bai, and S. Sclaroff, "Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 5851–5860.
- [272] H. Zhang, J. Wang, J. Zhang, T. Zhang, and B. Zhong, "One-stream vision-language memory network for object tracking," *IEEE Trans. Multimedia*, vol. 26, pp. 1720–1730, 2023.

- [273] J. Zuo, T. Wu, M. Shi, X. Liu, and X. Zhao, "Multi-modal object tracking with vision-language adaptive fusion and alignment," in *International Conference on Robotics, Intelligent Control and Artificial Intelligence*, 2023, pp. 1125–1133.
- [274] C. Zhang, X. Sun, Y. Yang, L. Liu, Q. Liu, X. Zhou, and Y. Wang, "All in one: Exploring unified vision-language tracking with multi-modal alignment," in *ACM Int. Conf. Multimedia*, 2023, pp. 5552–5561.
- [275] D. Ma and X. Wu, "Capsule-based object tracking with natural language specification," in *ACM Int. Conf. Multimedia*, 2021, pp. 1948–1956.
- [276] Y. Ma, Y. Tang, W. Yang, T. Zhang, J. Zhang, and M. Kang, "Unifying visual and vision-language tracking via contrastive learning," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4107–4116.
- [277] L. Zhou, Z. Zhou, K. Mao, and Z. He, "Joint visual grounding and tracking with natural language specification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 23 151–23 160.
- [278] Z. Yang, T. Kumar, T. Chen, J. Su, and J. Luo, "Grounding-tracking-integration," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 9, pp. 3433–3443, 2020.
- [279] Y. Shao, S. He, Q. Ye, Y. Feng, W. Luo, and J. Chen, "Context-aware integration of language and visual references for natural language tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 208–19 217.
- [280] R. Wang, Z. Tang, Q. Zhou, X. Liu, T. Hui, Q. Tan, and S. Liu, "Unified transformer with isomorphic branches for natural language tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 9, pp. 4529–4541, 2023.
- [281] H. Li, X. Liu, G. Li, S. Wang, L. Qing, and Q. Huang, "Boost tracking by natural language with prompt-guided grounding," *IEEE Trans. Intell. Transport. Syst.*, vol. 26, no. 1, pp. 1088–1100, 2025.
- [282] Y. Li, J. Yu, Z. Cai, and Y. Pan, "Cross-modal target retrieval for tracking by natural language," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 4931–4940.
- [283] Y. Sun, F. Yu, S. Chen, Y. Zhang, J. Huang, Y. Li, C. Li, and C. Wang, "Chattracker: Enhancing visual tracking performance via chatting with multimodal large language model," vol. 37, 2024, pp. 39 303–39 324.
- [284] H. Zhu, P. Zhang, L. Xue, and G. Yuan, "Multi-modal understanding and generation for object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 35, no. 5, pp. 4384–4396, 2025.
- [285] G. Zhang, B. Zhong, Q. Liang, Z. Mo, N. Li, and S. Song, "One-stream stepwise decreasing for vision-language tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 10, pp. 9053–9063, 2024.
- [286] Z. Mao, Y. Wang, X. Chen, and J. Yan, "Textual tokens classification for multi-modal alignment in vision-language tracking," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 8025–8029.
- [287] X. Li, S. Hu, X. Feng, D. Zhang, M. Wu, J. Zhang, and K. Huang, "How texts help? a fine-grained evaluation to reveal the role of language in vision-language tracking," *arXiv preprint arXiv:2411.15600*, 2024.
- [288] M. Guo, Z. Zhang, L. Jing, H. Ling, and H. Fan, "Divert more attention to vision-language object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8600–8618, 2024.
- [289] M. Guo, Z. Zhang, H. Fan, and L. Jing, "Divert more attention to vision-language tracking," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 4446–4460, 2022.
- [290] L. Shi, B. Zhong, Q. Liang, X. Hu, Z. Mo, and S. Song, "Mamba adapter: Efficient multi-modal fusion for vision-language tracking," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.
- [291] H. Zhao, X. Wang, D. Wang, H. Lu, and X. Ruan, "Transformer vision-language tracking via proxy token guided cross-modal fusion," *Pattern Recognit. Lett.*, vol. 168, pp. 10–16, 2023.
- [292] J. Ge, J. Cao, X. Zhu, X. Zhang, C. Liu, K. Wang, and B. Liu, "Consistencies are all you need for semi-supervised vision-language tracking," in *ACM Int. Conf. Multimedia*, 2024, pp. 1895–1904.
- [293] C. Zhang, L. Liu, H. Wen, X. Zhou, and Y. Wang, "Mambatrack: Exploiting dual-enhancement for night uav tracking," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025, pp. 1–5.
- [294] J. Yang, Z. Li, F. Zheng, A. Leonadis, and J. Song, "Prompting for multi-modal tracking," in *ACM Int. Conf. Multimedia*, 2022, pp. 3492–3500.
- [295] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [296] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9516–9526.
- [297] Z. Li, W. Cai, J. Dong, J. Lai, and X. Xie, "Feature disentanglement and adaptive fusion for improving multi-modal tracking," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 68–80.
- [298] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [299] S. Gong, Z. Teng, R. Li, J. Fan, B. Zhang, and J. Fan, "Minet: Modality interaction network for unified multi-modal tracking," *Image Vis. Comput.*, vol. 148, p. 105071, 2024.
- [300] M. Sun, X. Liu, H. Wang, and J. Liu, "Mixrgbx: Universal multi-modal tracking with symmetric mixed attention," *Neurocomputing*, p. 128274, 2024.
- [301] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 13 608–13 618.
- [302] P. Ye, G. Xiao, and J. Liu, "Amatrack: A unified network with asymmetric multimodal mixed attention for rgbd tracking," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2024.
- [303] Y. He, Z. Ma, X. Wei, and Y. Gong, "Knowledge synergy learning for multi-modal tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 7, pp. 5519–5532, 2024.
- [304] Z. Tang, T. Xu, X. Wu, X.-F. Zhu, and J. Kittler, "Generative-based fusion mechanism for multi-modal tracking," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 6, 2024, pp. 5189–5197.
- [305] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen *et al.*, "Onetracker: Unifying visual object tracking with foundation models and efficient tuning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 079–19 091.
- [306] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu *et al.*, "Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 26 551–26 561.
- [307] X. Chen, B. Kang, J. Zhu, D. Wang, H. Peng, and H. Lu, "Unified sequence-to-sequence learning for single- and multi-modal visual object tracking," 2024.
- [308] X. Hu, Y. Tai, X. Zhao, C. Zhao, Z. Zhang, J. Li, B. Zhong, and J. Yang, "Exploiting multimodal spatial-temporal patterns for video object tracking," in *AAAI Conf. Artif. Intell.*, vol. 39, no. 4, 2025, pp. 3581–3589.
- [309] T. Zhang, Q. Zhang, K. DeBattista, and J. Han, "Cross-modality distillation for multi-modal tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2025.
- [310] Y. Li, B. Wang, and Y. Li, "Lightfc-x: Lightweight convolutional tracker for rgb-x tracking," *arXiv preprint arXiv:2502.18143*, 2025.
- [311] C. Liu, Z. Guan, S. Lai, Y. Liu, H. Lu, and D. Wang, "Emtrack: Efficient multimodal object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 35, no. 3, pp. 2202–2214, 2025.
- [312] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D. P. Paudel, L. Van Gool, and R. Timofte, "Single-model and any-modality for video object tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 156–19 166.
- [313] Y. Tan, Z. Wu, Y. Fu, Z. Zhou, G. Sun, C. Ma, D. P. Paudel, L. Van Gool, and R. Timofte, "Towards a generalist and blind rgb-x tracker," *arXiv preprint arXiv:2405.17773*, 2024.
- [314] Z. Tang, T. Xu, X.-J. Wu, and J. Kittler, "M3track: Meta-prompt for multi-modal tracking," *IEEE Signal Process. Lett.*, vol. 32, pp. 1705–1709, 2025.
- [315] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, 2021.
- [316] X. Hu, B. Zhong, Q. Liang, L. Shi, Z. Mo, Y. Tai, and J. Yang, "Adaptive perception for unified visual multi-modal object tracking," *IEEE Trans. Artif. Intell.*, pp. 1–11, 2025.
- [317] H. Wang, T. Xu, Z. Tang, X.-J. Wu, and J. Kittler, "Uastrack: A unified adaptive selection framework with modality-customization in single object tracking," *arXiv preprint arXiv:2502.18220*, 2025.
- [318] X. Chen, B. Kang, W. Geng, J. Zhu, Y. Liu, D. Wang, and H. Lu, "Sutrack: Towards simple and unified single object tracking," in *AAAI Conf. Artif. Intell.*, vol. 39, no. 2, 2025, pp. 2239–2247.
- [319] Y. Tian, L. Xie, J. Qiu, J. Jiao, Y. Wang, Q. Tian, and Q. Ye, "Fast-iptn: Integrally pre-trained transformer pyramid network with token migration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9766–9779, 2024.
- [320] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "An accelerated correlation filter tracker," *Pattern recognition*, vol. 102, p. 107172, 2020.

- [321] D. A. Klein, D. Schulz, S. Frintrop, and A. Cremers, “Bobot-bonn benchmark on tracking,” *Univ. of Bonn*, 2010.
- [322] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, “Event-based moving object detection and tracking,” pp. 1–9, 2018.
- [323] C. Tang, X. Wang, J. Huang, B. Jiang, L. Zhu, J. Zhang, Y. Wang, and Y. Tian, “Revisiting color-event based tracking: A unified network, dataset, and metric,” *arXiv preprint arXiv:2211.11010*, 2022.
- [324] X. Wang, J. Huang, S. Wang, C. Tang, B. Jiang, Y. Tian, J. Tang, and B. Luo, “Long-term frame-event visual tracking: Benchmark dataset and baseline,” *arXiv preprint arXiv:2403.05839*, 2024.
- [325] J. Zhang, B. Dong, Y. Fu, Y. Wang, X. Wei, B. Yin, and X. Yang, “A universal event-based plug-in module for visual object tracking in degraded conditions,” *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1857–1879, 2024.
- [326] Z. Li, R. Tao, E. Gavves, C. G. Snoek, and A. W. Smeulders, “Tracking by natural language specification,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6495–6503.
- [327] H. Fan, H. Bai, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, Harshit, M. Huang, J. Liu *et al.*, “Lasot: A high-quality large-scale single object tracking benchmark,” *Int. J. Comput. Vis.*, vol. 129, pp. 439–461, 2021.
- [328] X. Li, S. Hu, X. Feng, D. Zhang, M. Wu, J. Zhang, and K. Huang, “Visual language tracking with multi-modal interaction: A robust benchmark,” *arXiv preprint arXiv:2409.08887*, 2024.
- [329] C. Zhang, L. Liu, G. Huang, H. Wen, X. Zhou, and Y. Wang, “Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark,” *arXiv preprint arXiv:2405.19818*, 2024.
- [330] X. Li, S. Hu, X. Feng, D. Zhang, M. Wu, J. Zhang, and K. Huang, “Dtvlt: A multi-modal diverse text benchmark for visual language tracking based on llm,” *arXiv preprint arXiv:2410.02492*, 2024.
- [331] C. Zhang, G. Huang, L. Liu, S. Huang, Y. Yang, X. Wan, S. Ge, and D. Tao, “Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9186–9205, 2022.
- [332] X.-F. Zhu, T. Xu, Z. Liu, Z. Tang, X.-J. Wu, and J. Kittler, “Unimod1k: Towards a more universal large-scale dataset and benchmark for multi-modal learning,” *Int. J. Comput. Vis.*, pp. 1–16, 2024.
- [333] A. Lu, M. Wen, J. Wang, Y. Guo, C. Li, J. Tang, and B. Luo, “Towards general multimodal visual tracking,” *arXiv preprint arXiv:2503.11218*, 2025.
- [334] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, “Trackingnet: A large-scale dataset and benchmark for object tracking in the wild,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.
- [335] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, J. Kämäräinen, and M. Danelljan, “The eighth visual object tracking VOT2020 challenge results,” in *Int. Conf. Comput. Vis. Workshops*, vol. 12539, 2020, pp. 547–601.
- [336] L. Wang, X. Zhang, H. Su, and J. Zhu, “A comprehensive survey of continual learning: Theory, method and application,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, 2024.
- [337] T. Xu, Z. Kang, X. Zhu, and X.-J. Wu, “Learning adaptive spatio-temporal inference transformer for coarse-to-fine animal visual tracking: algorithm and benchmark,” *Int. J. Comput. Vis.*, vol. 132, no. 7, pp. 2698–2712, 2024.
- [338] G. E. Karniadakis, I. G. Kevrekidis *et al.*, “Physics-informed machine learning,” *Nat. Rev. Phys.*, vol. 3, no. 6, pp. 422–440, 2021.
- [339] Y. Liu, Z. Liang, Y. Wang *et al.*, “Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 9087–9097.

Zhangyong Tang is now a Ph.D. student with the School of Internet of Things Engineering, Jiangnan University. His research interests include multi-modal object tracking and information fusion. He has published several scientific papers, including IJCV, TOMM, CVPR, AAAI, ACMMM, etc. He achieved top 3 performance in several competitions, including the VOT2020 RGBT challenge (ECCV2020), VOT2021 D challenge (ICCV2021), VOT2022 RGBD challenge (ECCV2022), and the 2nd Anti-UAV challenge.

Tianyang Xu (Member, IEEE) received his Ph.D. degree at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China, in 2019. He is currently an Associate Professor at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include visual tracking and deep learning. He has published several scientific papers, including IEEE TPAMI, IJCV, ICCV, IEEE TIP, IEEE TIFS, IEEE TKDE, CVPR, ECCV, ICCV, etc. His publications have been cited more than 5000 times. He achieved top 1 performance in several competitions, including the VOT2020 RGBT challenge (ECCV20), Anti-UAV challenge (CVPR20), and MMVRAC challenges (ICCV2021, ICME2024).

Xuefeng Zhu received the Ph.D. degree in Jiangnan University, Wuxi, China, in 2024. He is currently a lecture with the School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include visual tracking and deep learning.

Hui Li received the Ph.D. degree in Jiangnan University, Wuxi, China, in 2021. He is currently an assistant professor with the School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include image fusion and visual object tracking.

Shaochuan Zhao received the Ph.D. degree in Jiangnan University, Wuxi, China, in 2024. He is currently a post-doc with School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. His research interests include visual tracking.

Zhou Tao (Senior Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University in 2016. He is currently a Full Professor at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include machine learning, computer vision, AI in healthcare, and medical image analysis. He is an Associate Editor of IEEE TNNLS, IEEE TIP, IEEE TMI, and Pattern Recognition.

Xiao-Jun Wu received the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively. From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, where he was promoted to Professor. He has been with the School of Information Engineering, Jiangnan University, since 2006, where he is a Professor of Computer Science and Technology. He was a Visiting Researcher with the Centre for Vision, Speech, and Signal Processing, University of Surrey, U.K., from 2003 to 2004. He has published over 300 papers. His current research interests include pattern recognition, computer vision, and computational intelligence. He was a Fellow of the International Institute for Software Technology, United Nations University, from 1999 to 2000. He was a recipient of the Most Outstanding Postgraduate Award from the Nanjing University of Science and Technology. He is currently a Professor in artificial intelligent and pattern recognition at the Jiangnan University, Wuxi, China. He is IAPR/AAIA Fellow. His research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems.

Chunyang Cheng is currently a Ph.D student with the school of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include image fusion.

Josef Kittler (Life Member, IEEE) received the Ph.D. degree from the University of Cambridge, in 1974. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook *Pattern Recognition: A Statistical Approach* and over 700 scientific papers. His publications have been cited around 70,000 times (Google Scholar). He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of *Pattern Recognition Letters*, *Pattern Recognition and Artificial Intelligence*, *Pattern Analysis and Applications*. He also served as a member of the Editorial Board of *IEEE Transactions on Pattern Analysis and Machine Intelligence* during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.