Review and Analysis of RGBT Single Object Tracking Methods: A Fusion Perspective

ZHIHAO ZHANG, National Innovation Institute of Defense Technology, China
JUN WANG, National Innovation Institute of Defense Technology, China
ZHULI ZANG, Beijing Normal University, China
LEI JIN, Beijing University of Posts and Telecommunications, China
SHENGJIE LI, Beijing University of Posts and Telecommunications, China
HAO WU*, School of Artificial Intelligence, Beijing Normal University, China
JIAN ZHAO*, Northwestern Polytechnical University & China Telecom AI Institute, China
ZHANG BO*, National Innovation Institute of Defense Technolog, China

Visual tracking is a fundamental task in computer vision with significant practical applications in various domains, including surveillance, security, robotics, and human-computer interaction. However, it may face limitations in visible light data, such as low-light environments, occlusion, and camouflage, which can significantly reduce its accuracy. To cope with these challenges, researchers have explored the potential of combining the visible and infrared modalities to improve tracking performance. By leveraging the complementary strengths of visible and infrared data, RGB-infrared fusion tracking has emerged as a promising approach to address these limitations and improve tracking accuracy in challenging scenarios. In this paper, we present a review on RGB-infrared fusion tracking. Specifically, we categorize existing RGBT tracking methods into four categories based on their underlying architectures, feature representations, and fusion strategies, namely feature decoupling based method, feature selecting based method, collaborative graph tracking method, and traditional fusion method. Furthermore, we provide a critical analysis of their strengths, limitations, representative methods, and future research directions. To further demonstrate the advantages and disadvantages of these methods, we present a review of publicly available RGBT tracking datasets and analyze the main results on public datasets. Moreover,we discuss some limitations in RGBT tracking at present and provide some opportunities and future directions for RGBT visual tracking, such as dataset diversity, unsupervised and weakly supervised applications. In conclusion, our survey aims to serve as a useful resource for researchers and practitioners interested in the emerging field of RGBT tracking, and to promote further progress and innovation in this area.

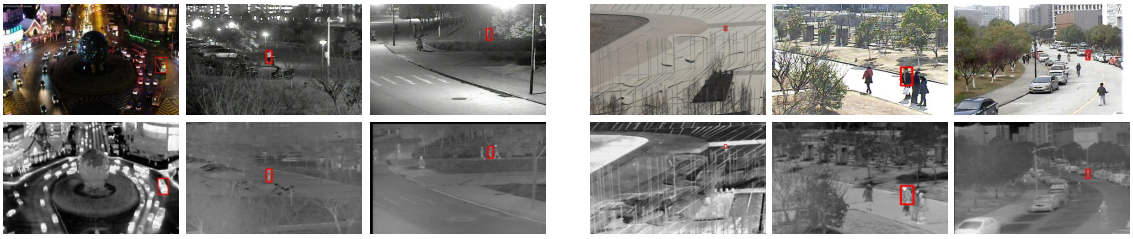
CCS Concepts: • **Computing methodologies** → **Tracking**.

Additional Key Words and Phrases: information fusion, RGBT, visual tracking, deep learning

*Corresponding authors.

Authors' addresses: ZhiHao Zhang, daerpq@outlook.com, National Innovation Institute of Defense Technology, Beijing, China; Jun Wang, National Innovation Institute of Defense Technology, Beijing, China; Zhuli Zang, Beijing Normal University, Beijing, China; Lei Jin, Beijing University of Posts and Telecommunications, Beijing, China; Shengjie Li, Beijing University of Posts and Telecommunications, Beijing, China; Hao Wu, School of Artificial Intelligence, Beijing Normal University, Beijing, China, wuhao@bnu.edu.cn; Jian Zhao, Northwestern Polytechnical University and & China Telecom AI Institute, Beijing, China, jian_zhao@nwpu.edu.cn; Zhang Bo, National Innovation Institute of Defense Technolog, Beijing, China, zhangbo10@nudt.edu.cn.

Manuscript submitted to ACM



(a) In nighttime tracking scenarios, infrared modality has more advantages over visible light

(b) In optimal lighting conditions with rich color features, visible light exhibits an advantage over infrared modalities.

Fig. 1. Examples of complementary information from visible and infrared images.

ACM Reference Format:
ZhiHao Zhang, Jun Wang, Zhuli Zang, Lei Jin, Shengjie Li, Hao Wu, Jian Zhao, and Zhang Bo. 2023. Review and Analysis of RGBT Single Object Tracking Methods: A Fusion Perspective. 1, 1 (March 2023), 27 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Single Object tracking aims to locate an object across a series of video frames [73, 92]. It plays an indispensable role in developing intelligent systems that can perceive and understand the world around them. Many emerging technologies rely on the ability to detect, track, and analyze moving objects in dynamic environments. For example, self-driving cars depend on tracking other vehicles, pedestrians, cyclists, and obstacles to navigate safely.

In literature, visible light tracking techniques are the predominant approaches in the field of object tracking. These methods utilize visual sensors like cameras to detect and follow the trajectory of moving objects in a scene. They are non-invasive, low-cost, and work in a variety of environments. However, the efficacy of computer vision systems predicated solely on visible light imaging is often constrained by various factors, encompassing occlusion, changes in illumination, the presence of cluttered backgrounds, and variations in object appearance. On the contrary, the utilization of infrared images provides significant complementary information for RGB tracking, encompassing superior immunity to variations in illumination, the capability to track objects even when partially or fully occluded, and the provision of more precise depth information. However, infrared data has less color information and indistinct features, which can make it unreliable in certain scenarios. Fig. 1 shows several examples of complementary between infrared and visible modality. To mitigate this gap, more researchers [48, 50, 64, 81, 102, 104, 112] pay attention to fusion algorithm to to overcome the tracking limitations of single-modal approaches.

Before the rise of deep learning, tracking methods were primarily based on traditional filtering techniques such as Bayesian filtering, particle filtering, mean shift, Kalman filtering, and correlation filtering. They often depend on manually designed features such as HOG [13], SIFT [62], and local binary pattern. The main criterion is that these methods use classical filtering techniques with a solid mathematical foundation. Feature decoupling based methods aim to decouple the feature representations from different modalities and learn the complex appearance representation of the object by modeling the unique characteristics of each modality separately. The effectiveness of these methods is measured by how well they decouple feature representations from the two modalities to enhance tracking performance.Feature selecting based methods focus on mining the discrimination information of different modalities for fusion. The criterion is how these methods employ unique strategies or generate weights related to modality in feature selection to optimize the tracking system's performance. Collaborative graph tracking methods divide the whole image into non-overlapping

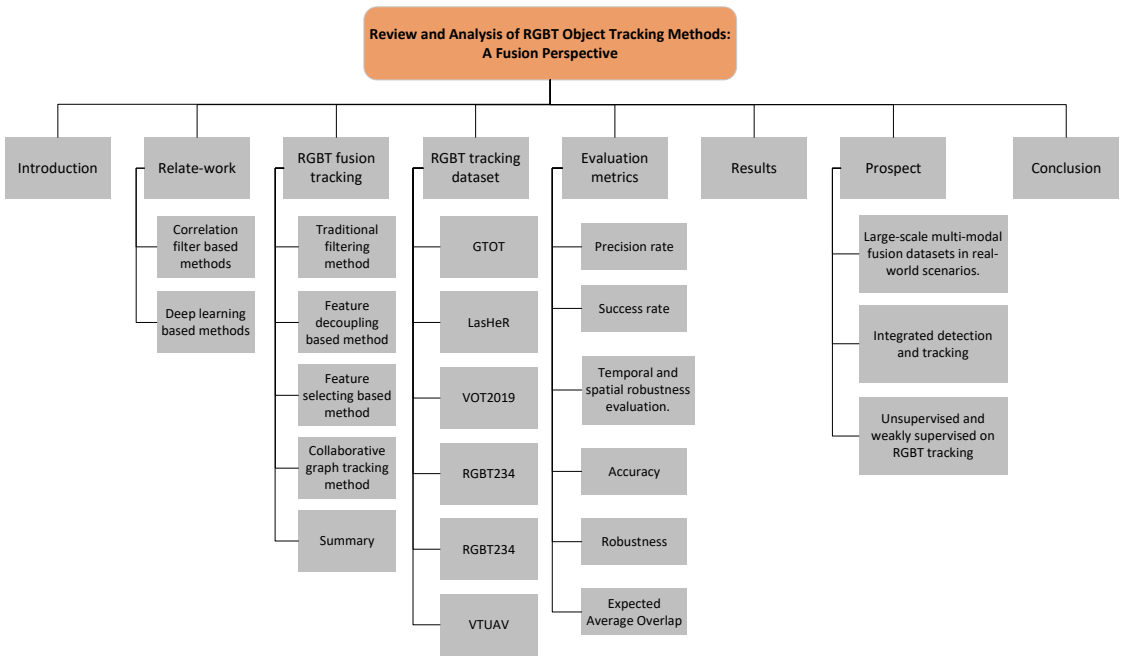


Fig. 2. Structure of this review.

image blocks and then construct the connection relationship in the form of a graph. The effectiveness of these methods is determined by how well they utilize graph models for RGBT target tracking.

In recent years, many works have been done to organize and review the methods for RGBT visual tracking. For example, Zhang *et al.* [106] provided a comprehensive review of RGBT tracking and systematically categorized target tracking methods based on different fusion methods and technical principles. Zhang *et al.* [101] sort out the RGBT object tracking framework according to the auxiliary function of different modalities. Xu *et al.* [75] conducted a comprehensive review of deep learning-based RGBT object tracking methods and identified differences in network architecture, categorizing them into two groups: MDnet-based [69] methods and Siamese network methods. However, the key to RGBT tracking lies in how to effectively fuse the information from multiple modalities, and existing surveys lack a comprehensive review of RGBT object tracking from the perspective of fusion. To enlighten further research on RGBT tracking, in this work, this paper will classify and analyze existing algorithms from the perspective of fusion. The main contributions of this review are in several aspects:

- We give a comprehensive overview of RGBT tracking, including benchmark datasets, performance metrics, and the systematic comparison of existing advanced methods.
- We summarize the performance of existing methods on some public benchmark datasets. Moreover, based on different technical principles, we conducted comprehensive analyses of different tracking methods, including their advantages, disadvantages, and representative methods in each category.
- We give detailed discussions on the future prospects and provide suggestions on promising research directions in this field.

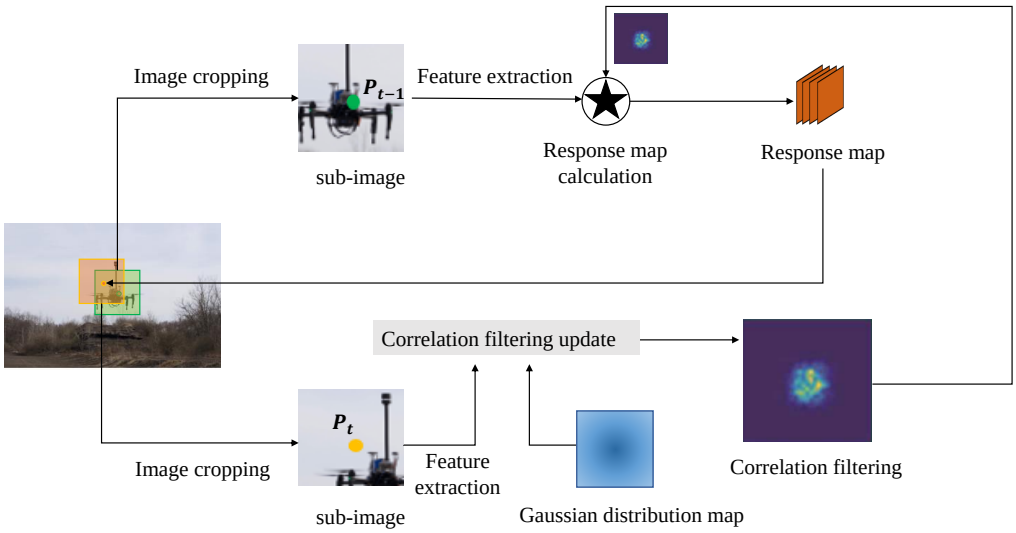


Fig. 3. Generic architecture of correlation filter trackers.

The structure of this review is schematically illustrated in Fig. 2. We briefly introduce the existing general trackers in Section 2. In Section 3, we discuss and analyze the RGBT trackers in detail. In Section 4 and Section 5, we present the related benchmark datasets and the performance metrics, respectively. Section 6 focuses on the experimental results and gives an analysis of the performances. In Section 7, we outline the future trends in RGBT tracking research. Finally, in Section 8, we conclude the paper.

2 RELATED WORKS

This section discusses some general object trackers that can aid in facilitating a deeper understanding of RGBT object tracking. It is noteworthy that this paper only briefly describes the RGB visual tracking method, which has a high correlation with RGBT visual tracking methods, such as correlation filter based methods and deep learning-based methods.

2.1 Correlation filter based methods

The main principle of the correlation filter is to calculate the correlation between two signals, as shown in Eqn. 1. To adapt to visual object tracking, the correlation filter based method aims to design a filter template and use this template to perform correlation operation with the target candidate region. The position of the maximum output response is the target position of the current frame. A simplified correlation based method diagram of the tracking pipeline is shown in Fig 3.

$$(u \otimes v)(\tau) = \int_{-\infty}^{+\infty} u^*(t)v(t+\tau)dt, \tag{1}$$

where u, v are two different signals, $*$ denotes complex conjugate.

Correlation filter tracking originated from the Minimum Output Sum of Squared Error (MOSSE) [4] proposed by David S. Bolme in 2010. It utilizes the correlation in signal processing, trains the correlation filter by extracting target features, and filters the input image of the next frame. CSK [31] extends the dense sampling and kernel trick on the basis of MOSSE. Dense sampling increases the number of samples without significantly increasing memory usage by circularly shifting the image vector like a rotating matrix. Kernel techniques can approximate high-dimensional space calculations in low-dimensional space and avoid dimensionality disasters. Danelljan *et al.* [17] embedded scale estimations into correlation filters to improve the tracking accuracy. [1, 22] added color features into tracking filters. Li *et al.* [61] incorporate both scale and color features for tracking. They design a set of hand-crafted features for further improvement, *i.e.*, HOG feature, color feature, and grayscale feature. Danelljan *et al.* [16] propose a target scale estimation algorithm in the tracking-by-detection framework. The algorithm utilizes a scale pyramid description to learn a Discriminative Correlation Filter (DCF) and trains two separate filters for scale estimation and position filtering, respectively. Based on this, SRDCF [19]further introduces penalty terms to eliminate boundary effects caused by image transformation.

However, manual features for correlation filter tracking can be limited in dealing with real challenges, particularly in complex environments. To overcome these limitations and improve tracking performance, there is an increasing trend toward utilizing deep learning techniques for feature extraction. [18] replace hand-crafted features with CNN features on the basis of the original SRDCF algorithm. C-COT [21] uses VGG-net to extract features and interpolates feature maps of different resolutions into a continuous spatial domain using the cubic spline function. Then, it uses the Hessian matrix to obtain the target position. To address the efficiency of tracking in terms of time and space, Matin *et al.* propose ECO [15] algorithm to optimize the C-COT algorithm on three levels: template updating, training dataset, and feature extraction.

The correlation filter algorithm is a ground-breaking technique in the field of object tracking and also a pioneering approach for RGBT tracking. While the existing correlation filtering algorithms may not achieve the same level of accuracy as deep learning based tracking algorithms, their main ideas have been consistently influencing the current state-of-the-art tracking algorithms.

2.2 Deep learning based methods

In complex tracking scenarios, traditional correlation filter methods may suffer from various challenging factors such as lighting changes, occlusions, scale variations, and non-rigid deformations, leading to a decline in tracking accuracy. In contrast, deep learning-based methods have shown stronger advantages in target feature extraction and modeling, thus exhibiting better robustness and accuracy in tracking tasks under complex environments. In this section, we summarize deep learning based single object tracking methods.

Siamese trackers The problem of target tracking involves determining the similarity between the search area and the template image. To address this problem, the Siamese network was introduced. This network shares a backbone network between the two branches, providing both with the same feature mapping ability. As shown in Fig. 4, a typical object siamese tracker follows a three-component framework, consisting of feature extraction, similarity measurement, and target position regression. Then, in this section, we will give a brief introduction to representative Siamese network based methods. Siamese visual trackers mostly utilize cross-correlation to measure the similarity between the search region and the template. For example, SiamFC [2], the pioneering work of siamese network, introduced the efficiency of end-to-end training scheme to the Siamese network. Due to its simple and effective structure, SiamFC has laid a solid foundation for the development of subsequent siamese network trackers. Therefore, many works have been extended

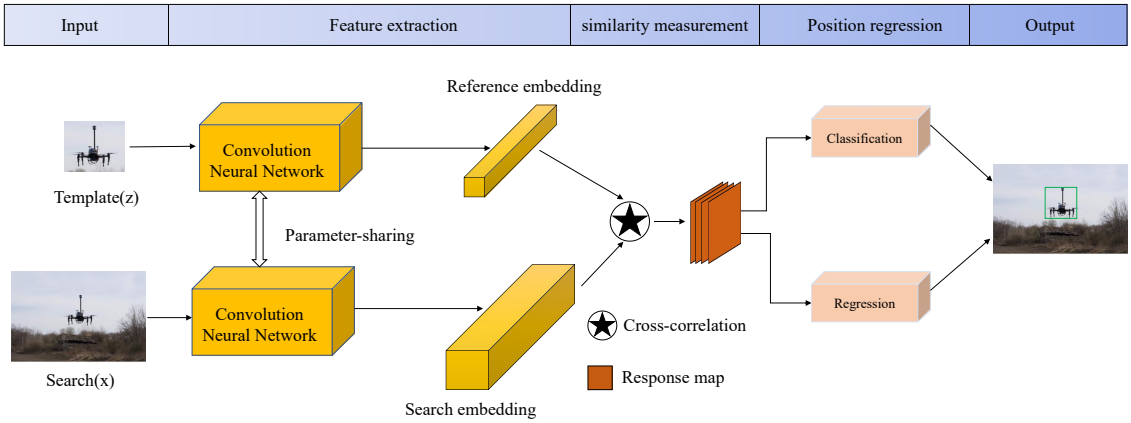


Fig. 4. Generic architecture of Siamese trackers.

based on SiamFC. Some of them have innovated from the perspective of network structure by adding new branches. For example,Li *et al.* [41] introduce RPN into Siamese network architecture and proposed SiamRPN. It divides the search area in the current frame into multiple grids, with each grid serving as a candidate box for RPN. The RPN network is then used to classify and regress these candidate boxes, ultimately obtaining the target's position. This work has greatly inspired researchers to explore the combination of RPN and siamese network architecture, resulting in a series of excellent trackers such as SiamRPN++ [39], C-RPN [23], SiamR-CNN [77], GlobalTrack [33], DaSiamRPN [114],*etc.* However, RPN-based trackers face difficulties in handling significant scale variations due to the need for the manual setting of anchor boxes. This limitation has led researchers to explore anchor-free network frameworks, which do not require prior knowledge of the number or distribution of anchor boxes.

SiamBAN [35] regards the visual tracking problem as a parallel classification and regression problem, directly classifying targets and regressing their bounding boxes in a unified, fully convolutional network, predefined anchors are removed so that the model parameters are reduced, and the speed is improved. The overall thinking of SiamCAR [29] and SiamBAN is consistent, but unlike SiamBAN, SiamCAR uses a splicing method to fuse the three extracted feature maps. This fusion method can continuously modify the network weights through later training to find the best fusion method. Compared to the direct weighted average fusion method of SiamBAN, this learnable fusion method has higher accuracy but may affect the speed.

To fully exploit the benefits of target localization classification and siamese network tracking methods, recent works such as the DIMP algorithm [3] and ATOM algorithm [14] have incorporated an IOU predictor into the siamese network architecture for scale estimation. Subsequently, an online-trained classifier is utilized for target localization. This approach allows for more accurate and efficient tracking performance in various applications.

Online trackers In actual tracking scenarios, the characteristics of the target (such as posture, appearance, texture, *etc.*) are prone to change. To solve this problem, researchers have added an online template update mechanism to the original tracking network, dynamically changing template features during the tracking process. In some initial works, simple linear interpolation is often used to update the template in each frame. However, this update mechanism needs to cope better with the challenges of deformation, occlusion, and fast motion in the tracking process. In order to solve these problems, some works take historical information into account, Danelljan *et al.* [20, 21] included a subset of historic frames as training samples which achieves better results than updating frame by frame, but increases the

amount of computation and memory usage. To accommodate this problem, Yang *et al.* [90] establish a template library and use LSTM to estimate the current template. Choi *et al.* [7]also establish a template library but use reinforcement learning to select the current template. On the basis of these works, Zhang *et al.* [100] propose UpdateNet, which learns the target features of the template from the first frame, the cumulative template, and the current frame respectively, to complete the template update, but it only uses historical information and does not consider layering feature. Some works perform updates based on gradient information. [74, 79] update the model by exploring the discriminative information in the backward gradient through multiple iterations. However, as the number of iterations increases, the real-time performance of the tracker will also decrease. To solve this problem, Li *et al.* [58] propose GradNet, which only needs one backpropagation and two forward passes to achieve template update. There are also some other works that have achieved good results. some meta-learning based methods [8, 12, 40, 109] focus on solving the problem of how and when templates are updated efficiently. DIMP [3] introduces an online learning strategy, which uses the target information in the tracking results to update online, while algorithms such as Dsiam [30], MDNet [69], and ATOM [14] achieve update effects by iteratively modifying model weights.

Transformer trackers With the development of transformer technology, more and more efforts have been made to introduce the encode-decode architecture into target tracking. For example, the deformable twin attention network SiamAttn [93] improves the network's feature expression ability by learning rich contextual information through a self-attention mechanism. Through a mutual attention mechanism, it interacts with information between templates and search areas before cross-correlation operations. In 2021, Chen *et al.* [6] propose a feature fusion network called TransT for attention mechanisms, using self-attention to enhance feature expression and then performing feature fusion through two cross-attention, taking the lead in replacing traditional cross-correlation operations with attention based feature fusion. TMT [57] and STARK [67] also have achieved good performance by introducing attention mechanism and feature fusion, but the transformer's global self-attention perspective will lead to over focusing on secondary information, blurring the edge area between the foreground and background, thus reducing tracking performance. Fu *et al.* [25] propose a Siamese tracking framework based on the sparse transformer to solve this problem, which improves performance and reduces training time. In addition, other research based on transformer: Song *et al.* [91] propose a cross-window attention mechanism to ensure the integrity of tracking objects. Cui *et al.* [82] propose an end-to-end framework based on an iterative hybrid attention module, which unifies feature extraction and target integration and has remarkable performance in short-term tracking.

3 RGBT FUSION TRACKING

In this paper, our approach to RGBT visual tracking centers on information fusion. Instead of solely adopting traditional methods and deep learning methods as classification criteria, we categorize existing RGBT visual tracking algorithms into four classes based on their approaches to multi-modal fusion, as illustrated in Fig. 5. In this section, we will deliver a thorough introduction encompassing technical principles and relevant algorithms for each category.

3.1 Traditional Filtering method

Since as early as 2000, the RGBT fusion approach has attracted researchers' attention. At that time, scholars widely used manual features such as HOG, SIFT, and local binary pattern (LBP) to process images of different modalities. In terms of tracking techniques, traditional tracking methods such as Kalman filtering, particle filtering, mean shift, and correlation filtering were mainly used for tracking.

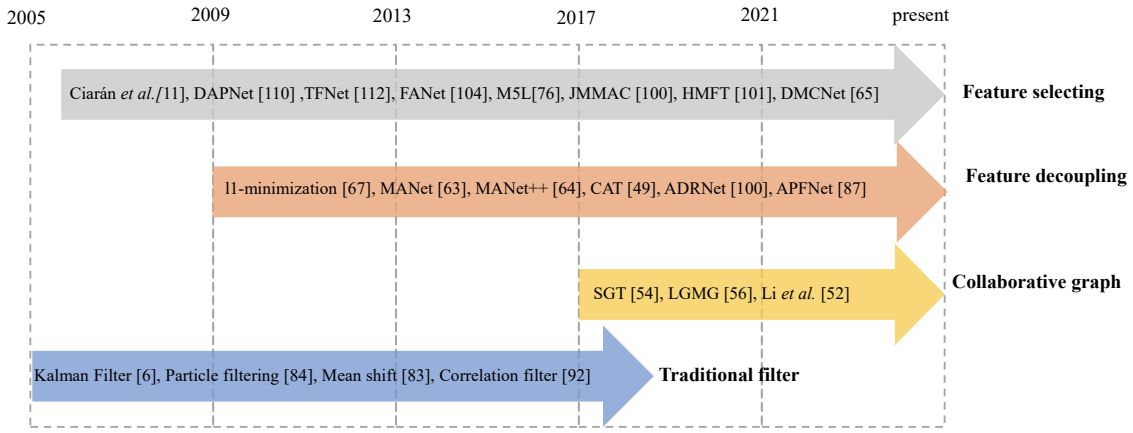


Fig. 5. We summarize the development of the RGBT tracking method in 18 years and list some typical algorithms.

The Kalman Filter is a widely used algorithm in object tracking that iteratively estimates the next moment's state by combining historical and observational data. Specifically, it models the motion of the target using a state-space model and incorporates measurements to improve the estimate of the target's position in the next frame. For example, Bunyak *et al.* [5] borrow the level set and combine it with the classic active contour method to propose a tracking method that fuses infrared and visible video. Yun *et al.* [86] proposes a flexible compressive time-space Kalman fusion tracking algorithm. It applies the compression tracking algorithm [98] to solve RGBT object tracking problem. However, the Kalman filter method is designed to address linear problems. Therefore, it may face substantial limitations in tracking scenarios that involve non-linear object motion, as commonly observed in practical applications.

Particle filtering is a Bayesian-based filtering method that estimates the state of a system using a set of weighted particles. Unlike the Kalman filter, particle filtering can handle non-linear and non-Gaussian systems. This makes it a versatile approach that can be applied to any form of state space model, regardless of the system's linearity. In 1998, Isard *et al.* [34] first introduced particle filter into object tracking. Several works have been done on this bias to perform RGBT object tracking. Cvejic *et al.* [11] conduct research on pixel-level fusion of visible and infrared images via performing particle filtering method. Moreover, experiments show performance differences between different modalities. Peteri *et al.* present a joint tracking method [70], which focuses on color features in visible light and temperature features in infrared modality, respectively. Xiao *et al.* [84] adopt a decision fusion mechanism to perform RGBT tracking while updating template information based on the fusion results. Besides, it assigns color weights to each particle and uses position information and color histograms to model the target.

Mean shift regards the feature space as a priori probability density function, and the input is treated as a set of sample points that satisfy a certain probability distribution. Xiao *et al.* [83] present a multi-cue mean-shift tracking (MMT) algorithm to perform RGBT tracking. As shown in Fig. 6, it uses a similarity-weighted algorithm to merge the multi-cue included gray histogram feature, the edge histogram feature, and the combined feature. Conaire *et al.* [10] propose a framework that can efficiently combine features for robust tracking based on fusing the outputs of multiple spatiogram trackers. This approach provides the flexibility to add, remove, or dynamically weight features. However, the mean shift algorithm is a local search algorithm prone to fall into local optimal solutions. In addition, the window size remains the same, lacking the ability to adjust the size of the prediction box dynamically.

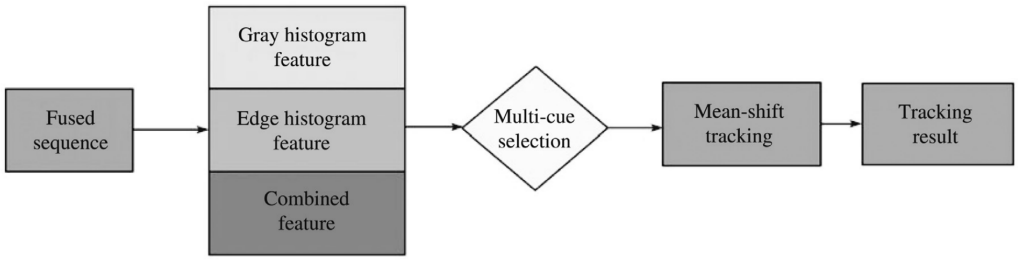


Fig. 6. The multi-cue mean-shift tracking algorithm.

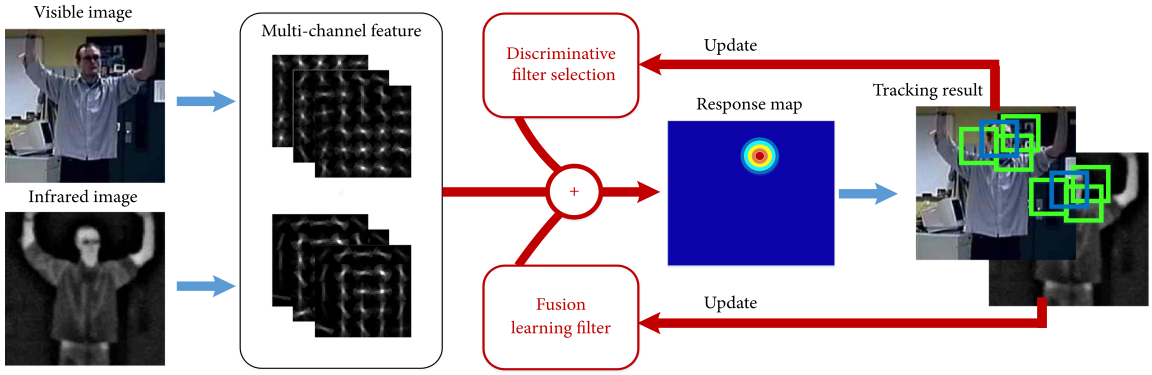


Fig. 7. The flowchart of discriminative fusion correlation learning model [94].

As mentioned in 2.1, the correlation filter is also an excellent tracking method due to its performance and high efficiency. To be best of our knowledge, Wang *et al.* [80] first apply CF-based fusion to RGBT tracking work. They take both collaboration and heterogeneity into consideration and propose a soft-consistent correlation filter (SCCF) for joint learning of the correlation filters of RGB and thermal spectra. Besides, they propose a weighted fusion mechanism to combine the response maps of the two modalities in order to compute the final response map in the detection phase. Zhai *et al.* [95] propose a fast RGBT tracking method via cross-modal correlation filters and also perform fusion in the final response map. This method sacrifices the reliability judgment of the modalities but achieves a very high tracking speed, reaching the fastest 224 FPS currently. Yun *et al.* propose a discriminative fusion correlation learning model (DFCL) [94], as shown in Fig. 7. In this work, given multichannel features from two modalities, the proposed discriminative filter selection and the fusion filter learning are applied to get the fusion response map. And then, the filters are updated via the tracking result obtained by the response map.

In summary, the early development of RGBT target tracking heavily relied on traditional methods, which provided a basis for future fusion approaches. While simple to implement and structure, these methods have limitations, such as using hand-crafted features that may need to provide more contextual information compared to deep features. Furthermore, many traditional methods are computationally intensive, posing difficulties in real-time implementation, except for correlation filtering methods. Thus, it is imperative to explore more advanced approaches that can effectively handle the challenges of RGBT target tracking, including the need for precise and efficient tracking, without being constrained by the limitations of traditional methods.

3.2 Feature Decoupling Based Method

Feature decoupling has been widely recognized as an effective approach to characterizing visual objects, including face recognition. In the context of RGBT tracking, the principle of feature decoupling is to learn separate feature representations for the target modality and subsequently combine them to enhance the overall tracking performance. The feature decoupling-based method excels in its strong representation capability for targets, achieving efficient encoding of target features with a minimal number of parameters. However, this approach has high requirements for the datasets, necessitating thorough feature annotations to complete the training process. These method may face difficulties in generalizing well across a wide range of tracking scenarios, as its effectiveness relies on the assumption that disentangling features enhances adaptability.

Several researchers have taken inspiration from this concept and proposed a range of popular algorithms, including sparse representation based feature decoupling fusion, attribute-based feature decoupling fusion, modality sharing, and specific feature decoupling fusion.

3.2.1 Sparse representation based feature decoupling fusion. Sparse representation is a mathematical technique that represents images as a combination of a few basic building blocks or atoms. The goal is to find a sparse set of coefficients that can represent the signal or image accurately. Mei *et al.* combine particle filter with sparse representation used in face recognition and propose the l1-minimization tracker [68], which lays the foundation of using sparse coding in visual tracking research. Subsequently, Wu *et al.* [81] first apply sparse representation to RGBT fusion tracking. The features of visible and infrared image patches are concatenated into a one-dimensional vector and then sparsely represented in the target template space. Liu *et al.* propose a joint sparse presentation approach which is utilized to design a fusion tracking approach for color and infrared images. In addition to directly using sparse representation for fusion, it can also be combined with various frameworks. Such as Li *et al.* [43] propose a real-time online grayscale-thermal tracking method via Laplacian sparse representation in the Bayesian filtering framework.

Considering the effect of modal reliability on tracking performance, Li *et al.* [42] propose a collaborative algorithm for online tracking. In this work, a reliability weight is introduced for each modality to realize the adaptive fusion. To increase the reliability of modal weight calculation, Li *et al.* [49] also use Laplacian sparse representation to design a multi-task model that exploits the similarity between image patches to optimize sparse coefficients. Lan *et al.* [38] present a feature representation and fusion model to combine the feature representation of the object in RGB and infrared modalities for object tracking. It performs feature representation of objects in different modalities by employing the robustness of sparse representation and combines the representation by exploiting the modality correlation.

Although sparse representation of RGBT tracking is innovative and practical, especially in suppressing characteristic noise, most of these algorithms can not meet the real-time requirements due to the time-consuming online optimization ratio of the sparse representation model, and this kind of model is generally based on pixel feature representation, which has poor robustness to complex scenes and environments.

3.2.2 Attribute-based feature decoupling fusion. In the field of visual tracking, researchers are faced with the challenge of modeling target appearance under various difficult conditions, such as fast motion, scale variation, illumination variation, *etc.* To address this challenge, researchers have focused on using a small number of parameters to represent complex target changes. Specifically, as shown in Fig. 8, many studies have attempted to decouple target characteristics based on the specific challenges that need to be addressed and have learned representations of corresponding attributes using attribute-based multi-branch networks.

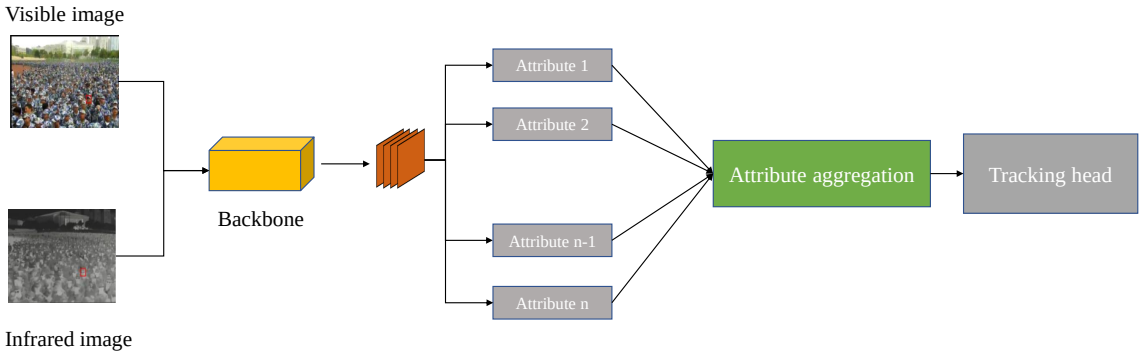


Fig. 8. The flow chart of attribute-based feature decoupling fusion strategy.

Qi *et al.* are the first to learn attribute-specific representations for visual tracking and propose an attribute-based neural network with multiple branches [71], where each branch is responsible for classifying the target under a specific attribute. By leveraging attribute information of video frames, this approach can generate a more discriminative representation, enabling it to tackle complex tracking challenges. Additionally, the design of the model is beneficial in reducing the appearance diversity of the target under each attribute, ultimately requiring fewer data to train the model. In consideration of the unique challenges that exist within each modality, as well as the common challenges that are shared between modalities, Li *et al.* have proposed the challenge-aware tracking algorithm (CAT) [48]. Specifically, this work considers illumination variation (IV) and thermal crossover (TC) as modality-specific challenges, while fast motion (FM), occlusion (OCC), and scale variation (SV) are treated as modality-shared challenges. In CAT, some branches aim to address the modality-sharing challenge via sharing parameters, while other branches with independent parameters focus on handling the modality-specific challenge.

The previous works mainly focus on a limited number of challenge factors, which may not be sufficient to cover all the possible challenges in real-world tracking scenarios. A natural extension is to design an adaptive attribute decoupling method based on the challenge factors. Such as Zhang *et al.* propose an attribute-driven representation network (ADRNet) [102], which considers not only major and special challenges such as occlusion and motion blur but also unknown coupled challenges. In particular, a general branch is designed to adaptively fit the attribute-agnostic tracking process. Xiao *et al.* propose an attribute-based progressive fusion network (APFNet) [87] by decoupling the fusion process of attribute information. APFNet adaptively aggregates attribute-specific fusion features using an aggregation model based on SKNet [60]. And the model can suppress noisy features from unappealing attributes by predicting the channel attention for each fusion feature.

Attribute-based feature decoupling effectively addresses various tracking challenges and can overcome them with limited training data. However, in practice, there may be diverse and unknown challenges. It is still necessary to explore how to design appropriate network structures to cope with unknown challenges and how to fully decouple the features to achieve a more accurate target representation.

3.2.3 Modality sharing and specific decoupling fusion. RGBT tracking aims to expand the data dimensions and achieve complementary fusion. However, the fusion process may result in data redundancy, making it challenging to distinguish valuable data from irrelevant or redundant data. Modality sharing and specific decoupling fusion have emerged as

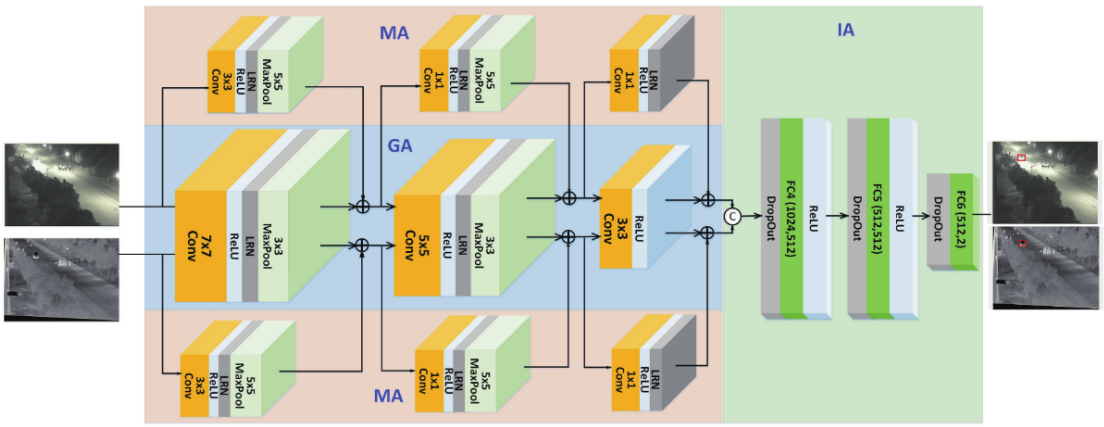


Fig. 9. Pipeline of MANet [64].

potential solutions to this problem, providing a promising approach for enhancing RGBT tracking. By sharing modality-specific features and decoupling modality-shared features, this approach can effectively reduce data redundancy and improve the representation of RGBT features. Therefore, it has gained significant attention in recent years as a way to improve the performance of RGBT tracking.

Li *et al.* propose the multi-adapter convolutional network (MANet) [64] to address the potential values of modality-shared cues and instance-aware information for RGBT tracking. MANet performs joint feature learning for modality-shared, modality-specific, and instance-aware representations in an end-to-end deep learning framework. Fig. 9 illustrates the three adapters used in MANet: the generality adapter extracts shared object representations, the modality adapter encodes modality-specific information to leverage their complementary advantages, and the instance adapter models the appearance properties and temporal variations of a specific object. The multi-adapter design of MANet improves the robustness and accuracy of RGBT tracking by exploiting the strengths of each modality and modeling instance-specific information. Based on MANet, Lu *et al.* propose MANet++ and hierarchical divergence loss [65]. Their loss function seeks to maximize the distribution difference between modality-specific and modality-shared features while minimizing the distance between the distribution of modality-shared features. In a similar vein, Peng *et al.* [36] propose SiamIVFN, a siamese infrared and visible light fusion network. Their approach involves a complementary feature fusion subnetwork, which uses filters with different coupling rates in each convolutional layer to learn the common features between infrared and visible images. Inspired by MANet, DMSTM [96] propose a dual-modality backbone. It performs elementwise addition between generic and modal features at each downsampling layer to achieve feature fusion at different scales, exploiting the spatial information at the shallow level with the semantic information at the deep level.

Modality sharing and specific decoupling fusion is a widely adopted technique to address the challenge of RGBT fusion. It simplifies the problem and enables accurate target modeling with limited parameters by eliminating redundant data. However, this approach has limited ability to complement the information between modalities because there is no interaction between the features of different modalities.

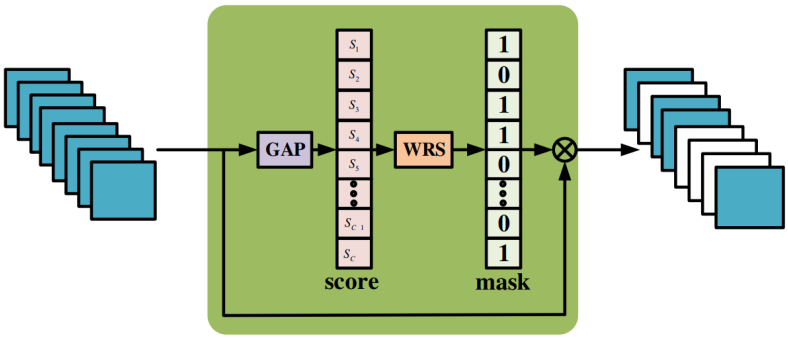


Fig. 10. Feature Pruning module [111] for selecting the channels with the highest score.

3.3 Feature Selecting Based Method

Feature selecting based methods are commonly used for RGBT fusion. They select the most informative and relevant features from each modality and fuse them together. This approach can effectively reduce the feature space's dimensionality, avoid data redundancy, and improve the model's generalization ability. According to different feature selection methods, Feature Selecting Based Methods can be divided into two categories: hard feature selection and soft feature selection.

3.3.1 Hard feature selection. Hard feature selection refers to selecting the most valuable features from the extracted features based on certain rules or criteria. And the selection process is usually done manually or using some pre-defined rules or algorithms.

In the early stages of RGBT object tracking, fixed weights with manual settings were used to integrate RGB and thermal features, as demonstrated in works such as [9, 10]. However, in these methods, the weights cannot adapt to quality variations of different modalities, resulting in suboptimal tracking performance. In 2016, Li *et al.* proposed an adaptive RGBT tracking method [42] in the Bayesian filtering framework. It introduces the weight variable for each modality and can optimize them online. Recently, deep learning-based hard feature selection methods have been applied to RGBT tracking. For example, Li *et al.* propose a two-stream fusion network in their work [50]. The network consists of a two-stream ConvNet and a FusionNet. The FusionNet is designed to fuse different modalities by adaptively selecting the most discriminative feature maps from the outputs of the two-stream ConvNet. During online tracking, the FusionNet is updated to ensure the best feature selection for adopting the appearance variation of the target. Following MDNet [69], Zhu *et al.* propose a dense feature aggregation and pruning network (DAPNet) [111]. As shown in Fig. 10, it employs the operations of global average pooling and weighted random selection algorithm to select the channels with the highest score. Furthermore, Zhu *et al.* present the trident fusion network (TFNet) [113], where the pruning strategy followed by DAPNet [111] is respectively applied to a single-modal branch and a multi-modal fusion branch.

The hard feature selection method has shown promising results in removing feature redundancy and noise. However, this approach heavily relies on manually designed loss functions or pruning criteria. Using hard selection may introduce risks to useful features, and mistakenly removing them could greatly reduce the accuracy of the algorithm.

3.3.2 Soft feature selection. Soft selection approaches calculate the weights of each modality based on their characteristics in order to achieve modality fusion. Compared to hard selection methods, soft selection can better adapt to the

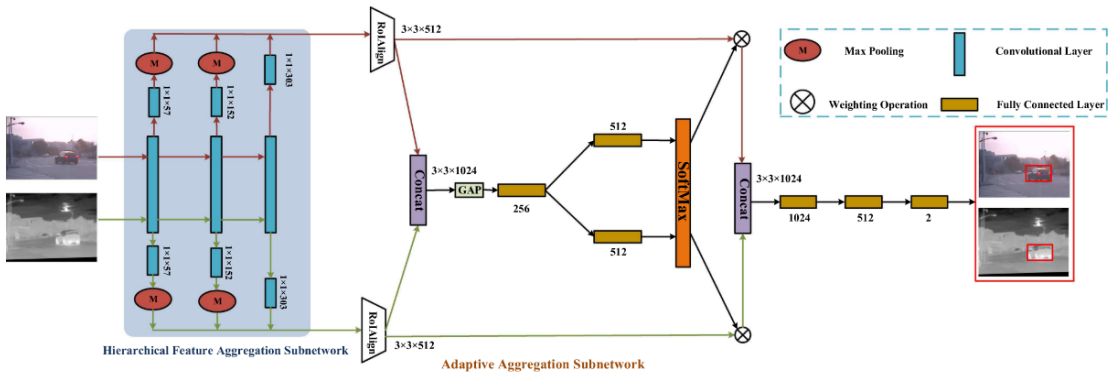


Fig. 11. The flowchart of FANet [112]

differences in feature distributions between different modalities, reduce the loss of feature information, and improve the effectiveness of modality fusion.

Built on the SiamFC [2] architecture, SiamFT [107] uses different fusion strategies for template and search features. Specifically, for template features, SiamFT uses simple concatenation, while for search features, the network learns modality reliability weights for fusion. Based on SiamFT, Zhang *et al.* proposed DSiamMFT [108], which is a fusion tracking method based on dynamic Siamese networks with multi-layer fusion. It employs an attention mechanism to calculate the reliability weight of multi-level features, enabling the adaptive fusion of multi-level and multi-modal features. Guo *et al.* propose a response-level fusion tracking method called DuSiamRT [27], which is based on a dual siamese network. It utilizes a modality-wise channel attention mechanism to evaluate the contribution of the channels of the two modal features. Built upon the TransT [6] architecture, Hou *et al.* propose a modality-aware tracker, termed MTNet [32]. In terms of specific fusion architecture, MTNet is similar to FANet. It construct a channel aggregation and distribution module to eliminate the redundant channels of backbone features. To obtain more accurate reliability weights of two modalities, Liu *et al.* propose Quality-Aware RGBT Tracker (QAT) [63] for robust RGBT tracking that combines reliability learning and residual guidance to enhance the features of each modality and improve tracking performance.

In contrast to the aforementioned feature-based soft selection Siamese network trackers, there exist several soft selection trackers based on other architecture, *i.e.* MDNet. For example, Zhu *et al.* propose a quality-aware feature aggregation network (FANet) [112]. As shown in Fig. 11, it first concatenates the features of two modalities for inter-modality interactions and then separate it from calculating the modality weights. Besides, it also takes the reliability of different layer features into consideration and proposes a hierarchical feature aggregation subnetwork for the adaptive aggregation of multi-layer depth features. Furthermore, Gao *et al.* propose DAFNet [26], a deep adaptive fusion network for RGBT tracking. Compared with FANnet, DAFNet adopts a progressive fusion framework and performs RGBT fusion on each layer of features during feature extraction. M5L [76] also adopts an attention-based fusion scheme to compute the importance of each modality. But they proposed Multi-modal Multi-margin Structured Loss from the perspective of positive and negative sample matching to preserve the structured information of samples. To further explore the potential of attention mechanisms in RGBT information fusion, some researchers have proposed to use of hybrid attention mechanisms to achieve this goal. For example, CBPNet [88] adopts channel attention and spatial attention mechanism to perform multi-modal cross-layer bilinear pooling tracking algorithm. JMMAC [103] divides the modality

weights into global weights and local weights to achieve more accurate fused response maps. Specifically, it uses global weights to exploit the complementarity of RGB and T modalities and obtains the weight over the whole context. The local weights, on the other hand, are used to suppress the influence of distractors from negative samples and improve the robustness of the tracker. Zhang *et al.* propose Hierarchical RGBT Fusion Tracker (HMFT) [104], which integrates fusion modules at three levels: image, feature, and decision. At the feature level, HMFT introduces a channel-wise modality weight to perform discriminative feature fusion. At the decision level, HMFT employs an Adaptive Decision Fusion (ADF) to these two response maps according to their modality confidences.To cope with changes in target size, the MSIFNet [85] algorithm design a feature selection module that adaptively selects multi-scale features for fusion by the channel-aware mechanism while suppressing noise and redundant information brought by multiple branches.

In some cases, traditional soft feature fusion methods using weighted fusion may suppress some useful information. Therefore, some researchers have attempted to enhance the interaction between multi-modal features to achieve bidirectional feature soft selection, with the dominant modality guiding the weaker modality. For example, in nighttime scenes, the infrared modality can be used to guide the feature representation of visible light modality rather than multiplying a small weight for the visible light modality. This interactive fusion method can improve the utilization efficiency of multi-modal features, thereby enhancing tracking performance. For example, Hui Lu *et al.* proposed an RGBT object tracking algorithm (MaCNet) [97] to effectively fuse dual-modality information. MaCNet uses the visual attention mechanism of each modality to estimate the importance of corresponding features and then guides the feature fusion process with shared features indicating modality importance to enhance interaction between modalities and improve overall tracking performance. Wang *et al.* propose a cross-modal pattern-propagation (CMPP) [78] tracking framework to diffuse instance patterns across RGBT data on the spatial domain as well as temporal domain. This work presents an inter-modal pattern propagation method, making useful patterns that may be mutually propagated between modalities so that feature information can be compensated for each other. Besides, it extends the spirit of pattern propagation from the cross-modal spatial domain to the temporal domain to construct dynamical pattern propagations. Zhang *et al.* propose a complementarity- and distractor-aware RGBT tracker (SiamCDA) [105]. It builds on an advanced anchor-based tracker, SiamRPN++ [39], and contains a complementarity-aware multi-modal feature fusion module (CA-MF) for multi-modal feature fusion before the Siamese region proposal procedure. The features of the two modalities are first to be enhanced by reducing modality differences between unimodal features. And then, enhanced RGB and thermal features will be further combined to achieve the final fused features via some fusion strategies. Similar to SiamCDA, SiamCAF [89] designed a Complementary Coupling Feature fusion module (CCF) to extract similar features and reduce modality differences to fuse features better. It first extracts the similar features between visible light and thermal infrared features using the coupled filter to obtain the weight maps. Then, it enhances the features using cross-modal connections. Finally, CCF fuses the enhanced features via concatenation and fuse the channel in the 1×1 convolutional layer to fuse the channel information. Lu *et al.* propose a duality-gated mutual condition network (DMCNet) [66] to exploit the discriminative information of all modalities while suppressing the effects of data noises. In this work, the duality-gated mutual conditional module is used to extract the modal discriminative features to guide the learning of the other modal features. Feng *et al.* propose a contribution-aware aggregation network [24] to adaptive learn the reliable weight of RGB and thermal modalities for fusion. The the fused response is passed through the classification and regression network to locate the target. Zhu *et al.* propose a visual prompt tracking framework [110].It considers visible light as the primary modality and infrared as the auxiliary modality, utilizing the supplemental information from infrared to better complement the primary modality.

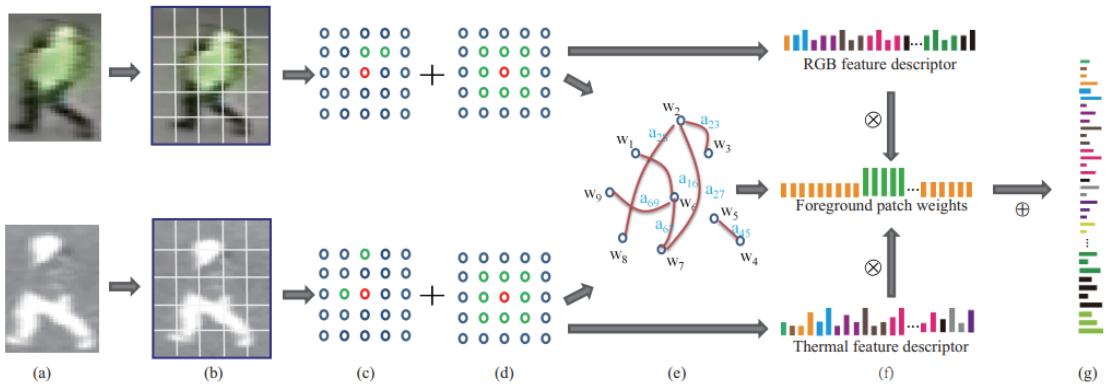


Fig. 12. A general framework of collaborative graph based RGBT tracking method.

A discriminative fusion method based on feature selection is one of the popular approaches currently, which has a natural advantage in modality redundancy reduction and modality reliability discrimination. With the development of attention mechanism in recent years, attention-based adaptive feature selection methods have become a research hotspot.

3.4 Collaborative graph tracking method

The collaborative graph tracking method is a classification-based tracking method that converts the object tracking problem into a binary classification problem. This method uses a classifier to distinguish between the target region and the background region. It uses a co-occurrence graph to capture the spatial structure information of the target region. Specifically, this method first segments the image and divide the sample region into a plurality of non-overlapping patches as nodes. Then this method assigns a weight to each image patch to indicate whether the patch belongs to the foreground or background. And the edge weights between two patches are also assigned to denote the relationship between them. Finally, a classification method such as SVM is used to classify the foreground and background among these nodes to obtain the object bounding box. Fig. 12 shows a general framework of the collaborative graph based RGBT tracking method.

Li *et al.* [53] first apply the graph model to RGBT tracking. This work proposes a weighted sparse representation regularized graph to jointly optimize modal weights, sparse representation coefficients, graph node similarity, and graph node weights. Furthermore, Li *et al.* design a cross-modal ranking algorithm [54] to compute the patch weights for suppressing background information. In 2022, Li *et al.* [51] separate the ranking process of different features (color and gradient features) and achieve more accurate weight calculation in a multi-task learning framework. In order to utilize local information and global information, Li *et al.* [55] propose an approach to learn a local-global multi-graph descriptor to suppress background effects for RGB-T tracking. It designs an optimization framework that can dynamically learn a joint graph with both local and global considerations using spatial smoothness and low-rank representation. Li *et al.* [56] also present a two-stage modality-graphs regularized manifold ranking for RGB-infrared tracking. In the first stage, the patch weight is computed based on the initial seeds. And at the second stage, the patch weight computation is based on the result of the first stage. Shen *et al.* [72] propose a cooperative low-rank graph model to suppress background clutter.

It decomposes input dual-modal features into low-rank components and sparse, noisy components and dynamically updates them by the collaborative graph learning algorithm.

Collaborative graph tracking methods have shown promising results in addressing challenges such as target variation and occlusion, and they can effectively suppress the background cluster through patch weight. This approach has been adopted in various general target tracking methods as well, such as those presented in [28, 46, 47]. However, most existing methods rely only on color and gradient features. Combining depth features with graph models is a potential future direction for improving the performance of collaborative graph tracking methods.

3.5 Summary

This paper provides a comprehensive review of existing RGBT tracking methods, which can be divided into four categories: feature decoupling based methods, feature selecting based methods, traditional methods, and collaborative graph tracking methods. In the early days of RGBT tracking, researchers mainly focused on using handcrafted features to discriminate targets and proposed methods such as sparse representation, Kalman filtering, particle filtering, mean shift, correlation filtering, and co-graph. These early works greatly promote the development of the RGBT object tracking field and provided a foundation for subsequent algorithms. With the development of deep learning, people begin to use deep features for modal fusion, and a large number of deep learning fusion algorithms have been proposed. One of the most important feature fusion methods is feature decoupling, such as using modal-shared and modal-specific information fusion, and using attribute-based feature decoupling fusion strategies, which were also proposed. However, these methods depend on the attribute labeling of training data and model optimization strategies. Currently, coupling features with different properties at different stages of feature extraction has become one of the research directions for this type of method.

Regarding another type of method, most feature soft selection based methods currently use attention mechanisms to determine the effectiveness of features to remove clutter. However, these methods ignore the enhancement of useful information. Currently, feature bidirectional soft selection methods based on attention mechanisms have begun to receive attention, which focuses on guiding another modality from one modality to achieve cross-fusion effects. With the continuous exploration of attention mechanisms, feature bidirectional selection based on attention mechanisms is gradually becoming a research hotspot.

4 RGBT TRACKING DATASETS

Large-scale datasets are crucial to the field of RGBT vision tracking, both for the training of deep learning algorithms and for the extensive evaluation of the performance of various trackers, thus effectively contributing to the research and development of the field.

This paper will present popular RGBT object tracking datasets, such as VOTRGBT, RGBT234, RGBT210, LasHeR, VTUAV, GTOT.

The first large-scale RGB-infrared dataset, GTOT[42], was presented in 2016, which included 50 pairs of recorded greyscale and infrared images videos under different scenarios and conditions and provided seven challenge attribute labels to evaluate the performance of the algorithm under different challenge attributes. It also contained annotation data, including bounding boxes around the target. To further enrich the diversity of RGBT visual target tracking datasets, Li *et al.* [53] proposed a larger scale RGBT visual tracking dataset, RGBT210, containing 210 pairs of RGBT video sequences. This dataset provided a total of approximately 210K frames, which was sufficient for a performance evaluation. However, the RGBT210 dataset is not sufficiently well labeled. To address this problem, Li *et al.* [45] improved

it in 2019 by proposing a larger RGBT tracking dataset, RGBT234, and providing 12 challenge attribute annotations to evaluate tracking performance in challenging situations. In the same year, the VOT committee selected 60 sequences from RGBT234 and constructed a new dataset, VOT2019RGBT [37], and used it in VOT-RGBT2019 and VOT-RGBT2020. The evaluation indicator for this dataset is Expected Average Overlap(EAO). Although the datasets mentioned above are large enough to evaluate the performance of different algorithms, they do not meet the need for large-scale training data for deep trackers. To address this issue, Li *et al.* [52] proposed LasHeR, the largest available RGBT tracking dataset, providing 1,224 pairs of RGBT video sequences with 19 challenge properties labeled, which will facilitate methodological research in the field of RGBT tracking. In addition, Zhang *et al.* [104] proposed an RGBT tracking dataset for unmanned aerial vehicle platforms, VTUAV, which contained 13 hierarchical attributes ,and they presented the problem of RGBT long-time tracking, which opens up a new research space in the RGBT field. Moreover, they provide a coarse-to-fine attribute annotation, where frame-level attributes are provided to exploit the potential of challenge-specific trackers. In order to understand the characteristics of different datasets more clearly, the details of the mainstream RGBT tracking datasets are summarized in Table 2.

Table 1. Common Challenge Attribute Set Abbreviation Comparison Table

Attribute	Description	Attribute	Description	Attribute	Description
NO	No Occlusion	DEF	Deformation	OCC	Partial or full occlusion
PO	Partial Occlusion	FM	Fast Motion	LSV	Large scale variation
HO	Hyaline Occlusion	SV	Scale Variation	SO	Small object
TO	Total Occlusion	ARC	Aspect Ratio Change	TB	Target Blur
LI	Low Illumination	MB	Motion Blur	EI	Extreme Illumination
HI	High Illumination	SA	Similar Appearance	FO	Full Occlusion
AIV	Abrupt Illumination Variation	OV	Out-of-View	TVS	Thermal-visible separation
LR	Low Resolution	CM	Camera Moving	TC	Thermal Crossover
BC	Background Clutter	FL	Frame Lost		

Table 2. Comparison of RGB-infrared fusion tracking datasets

Name	Videos	Frames (In total)	Attributes	Ground truth	Video type	Resolution	Year
GTOT	50	15.8K	7	Yes	Gray, T	Various	2016
RGBT210	210	210K	12	Yes	RGB, T	630 × 460	2017
RGBT234	234	234K	12	Yes	RGB, T	630 × 460	2019
VOT2019RGBT	60	40.2k	12	Yes	RGB, T	630 × 460	2019
LasHeR	1224	1224K	19	Yes	RGB, T	630 × 460	2021
VTUAV	500	1700K	13	Yes	RGB, T	1920 × 1080	2022

5 EVALUATION METRICS

In recent years, several well-recognized evaluation metrics have been proposed to evaluate the performance of visible image-based tracking. These metrics include Precision Rate (PR), Success Rate (SR), Accuracy, Rolodex, and Expected Average Overlap (EAO). These evaluation metrics can also be applied to RGBT fusion tracking.

Precision Rate. Precision measures the accuracy of predicted bounding boxes by assessing the position error (CLE) in relation to the ground truth. It is considered successful if the CLE is less than or equal to a specified threshold, often set at 20 pixels. Precision, depicted in a precision diagram, indicates the percentage of accurately predicted frames

under this threshold. However, precision overlooks target scale changes and may lead to increasing errors after target loss. Li *et al.* [45] proposed using Maximum Precision Rate (MPR) for fusion tracking assessment. MPR calculates the Euclidean distance between predicted and true bounding boxes in both visible and infrared modes, selecting the smaller distance for precision scoring.

Success rate The success rate represents the intersection ratio between the predicted box and ground truth bounding box is larger than a threshold. The overlap is defined as follows:

$$O(a, b) = \frac{A \cap B}{A \cup B}. \tag{2}$$

Li *et al.* [44] proposed to utilize the maximum success rate (MSR) to evaluate fusion tracking performance. Specifically, we calculate the overlaps ratio between the predicted bounding box and the true box in both visible and infrared modes for each frame. The larger overlap value is chosen to calculate the success score.

Temporal robustness evaluation and spatial robustness evaluation. The above two common one-pass evaluation (OPE) metrics have two drawbacks. One is that a tracking algorithm may be sensitive to the initial position given in the first frame, which can have a significant impact at different positions or initial frames. Secondly, most algorithms do not have a mechanism to reinitialize after encountering tracking failures. For these two issues, researchers propose Temporal Robustness Evaluation (TRE) and Spatial Robustness Evaluation (SRE).

For TRE, each tracking algorithm is evaluated numerous times from different starting frames across an image sequence. In each test, an algorithm is evaluated from a particular starting frame, with the initialization of the corresponding ground-truth object state, until the end of an image sequence. The tracking results of all the tests are averaged to generate the TRE score.

For SRE, each tracking algorithm will be initialized by slightly shifting or scaling the ground truth bounding box of a target object. The core idea is to evaluate whether a tracking method is sensitive to the initialization state by running several tests.

Accuracy. The accuracy measures how well the bounding box predicted by the tracker overlaps with the ground truth bounding box.

Robustness. Object tracking often fails to achieve overlap or overlap ratio with the ground truth bounding box due to various reasons, resulting in losing the target. A good tracking algorithm needs to prevent this situation from happening, and the measurement standard is robustness (R), which is defined as the proportion of times the tracker loses the target during the tracking period.

Expected Average Overlap. The EAO (Expected Average Overlap) is a comprehensive evaluation metric for short-term tracking in VOT. It balances the two aspects of Accuracy and Robustness to estimate how accurate the estimated bounding box is after a certain number of frames are processed since initialization.

6 RESULTS

In this section, we evaluate some representative algorithms on currently available public datasets, including the GTOT, RGBT234, LasHeR, RGBT210, and VTUAV.

GTOT. The GTOT dataset is one of the benchmark datasets for RGBT object tracking, and most RGBT trackers are evaluated based on their success rate and precision rate. Talbel 3 presents the performance results of 23 trackers on this dataset. As can be seen from Table 3, algorithms based on deep learning feature selection and feature decoupling have achieved promising results on this dataset, such as CMPP [78], DMCNet [66], APFNet [87], SiamIVFN [36], SiamCDA [105], *etc.*

RGBT210 and RGBT234. RGBT234 is also a large-scale RGBT tracking dataset, which is an extension of the RBT210 dataset. RGBT234 contains 234 pairs of visible and thermal videos and 12 annotated attributes. Table 3 clearly shows the tracking results of 22 presented trackers. QAT achieves the best success rate of 64.3%. And it should be noted that both RGBT210 and RGBT234 datasets use the metrics of MPR and MSR to evaluate trackers instead of PR and SR. We further investigate the attribute-based comparison of MPR/MSR of 6 trackers on RGBT234, as shown in Table 4.

LasHeR. For the short-term RGBT tracking dataset (LasHeR), ten trackers are presented in this paper. Table 3 shows the success rate and precision rate among these trackers. In this dataset, ViPT achieves the best Success rate (0.525). Furthermore, we select six trackers with better tracking performance and present their results under various challenging scenarios, as shown in Table 6.

VTUAV. This dataset was proposed in 2022, and there are currently few algorithms using it for evaluation. To our best knowledge, QAT achieves the best success rate (66.7%) and precision rate (80.1%).

VOT-RGBT. In 2019 and 2020, the VOT community held two competitions on RGBT tracking issues, respectively. In this benchmark, Accuracy, Robustness, and EAO are employed to evaluate trackers. As shown in Table 5, JMMAC ranks first on the public dataset twice.

Analysis. RGBT tracking methods have made significant progress in recent years. Especially with the development of deep learning, feature decoupling, and feature selection methods have achieved remarkable improvements in tracking accuracy and robustness. For example, achieving an admirable tracking success rate of 74.9% in the GTOT dataset, which may indicate this method can be applied to automated tracking tasks. However, there are still some problems in RGBT tracking.

One of the major challenges in RGBT tracking is handling difficult scenarios such as out-of-view and background clutter. As shown in Table 6, even one of the best-performing tracker (DMCNet) achieves only 31.7% accuracy in the out-of-view challenge of the LasHeR dataset. Therefore, improving tracking accuracy requires addressing various challenging factors. This can be achieved by establishing larger and more comprehensive datasets for algorithm learning and by combining RGBT tracking with other methods, such as long-term tracking methods that integrate detection and tracking. Furthermore, the performance of existing RGBT tracking algorithms varies significantly across different datasets, making it difficult to adapt to real-world environments with high variability. To address this issue, unsupervised and weakly supervised learning methods can be used to enable the network to better mine abundant information in samples and broaden the generalization of the model.

In the next section, we will address the aforementioned issues by focusing on three key perspectives: datasets, tracking frameworks, and learning mechanisms.

7 PROSPECT

Despite the remarkable progress that has been made in RGBT fusion tracking, there are still some things that need to be fixed for future work. In this section, we give detailed discussions on specific trends of RGB-infrared fusion tracking based on the review of existing approaches.

7.1 Large-scale RGBT fusion datasets in real-world scenarios.

In real-world scenarios, visible and infrared data are often unaligned due to the position of the cameras, and the input data is also more complex and diverse, full of uncertainties. For example, the visible light dataset is almost ineffective in nighttime tracking scenes. The background, infrared heat sources, and multiple challenging factors heavily interfered

Table 3. Quantitative results of the existing deep RGBT trackers on GTOT, RGBT210, RGBT234, LasHeR and VTUAV datasets. The red bold fonts and blue fonts indicate the best and the second best performance.

Method	Type	GTOT P/S	RGBT210 P/S	RGBT234 P/S	LasHeR P/S	VTUAV P/S	FPS
SCCF [80]	Traditional	85/68.1	-	-	-	-	50
SGT [53]	Graph	85.1/62.8	-	72.0/47.2	32.7/23.2	-	5
[54]	Graph	82.7/64.3	-	-	-	-	8
CSR [42]	Decoupling	75.0/62.0	-	46.3/32.8	-	-	1.6
SiamFT [108]	Decoupling	75.8/62.3	-	68.8/48.6	-	-	30
MANet [64]	Decoupling	89.4/72.4	-	77.7/53.9	45.5/32.6	-	1.11
MANet++ [65]	Decoupling	90.1/72.3	-	80.0/55.4	46.7/31.4	-	25
MaCNet [97]	Decoupling	88.0/71.4	-	79.0/55.4	48.3/35.2	-	1
CAT [48]	Decoupling	88.9/71.7	79.2/53.3	80.4/56.1	45.0/31.4	-	20
ADRNet [102]	Decoupling	90.4/73.9	77.8/53.4	80.9/57.1	-	62.2/46.6	25
APFNet [87]	Decoupling	90.5/73.7	-	82.7/57.9	50.0/36.2	-	-
SiamIVFN [36]	Decoupling	91.5/79.3	-	81.1/63.2	-	-	147
DMSTM [96]	Decoupling	92.9/75.9	-	78.6/56.2	-	-	27.6
[50]	Selecting	85.2/62.6	-	-	-	-	15
DuSiamRT [27]	Selecting	76.7/62.8	-	56.7/38.4	-	-	116
DAPNet [111]	Selecting	88.2/70.7	-	76.6/53.7	43.1/31.4	-	2
CBPNet [88]	Selecting	88.5/71.6	-	79.4/54.1	-	-	3.7
M5L[76]	Selecting	89.6/71	-	79.5/54.2	-	-	14
DSiamMFT [108]	Selecting	-	64.2/43.2	-	-	-	14
DAFNet [26]	Selecting	89.1/71.2	72.6/48.5	79.6/54.4	44.8/31.1	62.0/45.8	23
FANet [112]	Selecting	89.1/72.8	-	78.7/55.3	44.1/30.9	-	17
mfDiMP [99]	Selecting	83.6/69.7	78.6/55.5	78.5/55.9	44.7/34.3	67.3/55.4	10.3
TFNet [113]	Selecting	89.1/72.8	77.7/52.9	80.6/56.0	-	-	17
HMFT [104]	Selecting	91.2/74.9	78.6/53.5	78.8/56.8	-	75.8/62.7	30.2
SiamCDA [105]	Selecting	87.7/73.2	-	76.0/56.9	-	-	37
JMMAC [103]	Selecting	90.2/73.2	-	79.0/57.3	-	-	4
CMPP [78]	Selecting	92.6/73.8	-	82.3/57.5	-	-	1.5
DMCNet [66]	Selecting	90.9/73.3	-	83.9/59.3	49.0/35.5	-	2.24
QAT [63]	Selecting	91.5/75.5	86.8/61.9	88.4/64.3	64.2/50.1	80.1/66.7	22
MSIFNet [85]	Selecting	90.5/74.1	-	81.7/57.0	-	-	-
MTNet [32]	Selecting	93.5/76.0	-	85.0/61.9	-	60.8/47.4	-
SiamCAF [89]	Selecting	90.6/73.0	-	77.1/53.7	-	67.0/54.1	-
SiamMLAA [24]	Selecting	91.3/75.1	77.9/56.7	79.5/58.4	53.8/43.1	53.8/43.1	21.7
ViPT [110]	Selecting	-	-	83.5/61.7	65.1/52.5	-	-

with the input image. However, existing trackers require highly registered data as input to achieve multimodal fusion, and various challenging factors are not as complex as in real life. Therefore, it is necessary to propose a large-scale unaligned dataset that includes various challenging factors in real-life scenarios.

7.2 Integrated detection and tracking

Current object tracking models require manual annotation of the first frame when tracking a specific video. However, in real-world automated tracking scenarios, the annotation information for the first frame may be missing, such as in unmanned anti-drone missions. Therefore, designing an automated tracking method that reduces the dependency on the first frame annotation is a worthwhile exploration for the future. With the help of object detection, the tracker

Table 4. Comparison results of PR/SR scores(%) of different trackers under different challenges (refering to Table ?? for detailed information) on RGBT234. The red bold fonts and blue fonts indicate the best and the second best performance.

Method	SiamCDA [105]	JAMMC [103]	CMPP [78]	DMCNet [66]	SiamIVFN [36]	MTNet [32]
NO	88.4/66.4	93.2/69.4	95.6/67.8	92.3/ 67.1	86.0/68.5	91.0/67.8
PO	84.2/63.9	84.1/61.1	85.5/60.1	89.5/63.1	83.0/65.3	88.7/ 64.8
HO	66.2/48.7	67.7/48.3	73.2/50.3	74.5/52.1	77.3/58.9	78.6/56.3
LI	81.8/61.2	84.0/58.8	86.2/58.4	85.3/58.7	81.0/62.1	83.3/59.5
LR	70.9/49.9	77.1/51.7	86.5/57.1	85.4/57.9	73.9/55.4	80.4/55.4
TC	67.4/47.7	74.9/52.6	83.5/58.3	87.2/61.2	64.9/48.3	86.1/61.6
DEF	77.9/59.2	70.6/52.9	75.0/54.1	77.9/56.5	79.6/63.0	84.7/64.0
FM	61.4/45.3	61.0/41.7	78.6/50.8	80.0/52.4	67.0/48.3	79.2/58.0
SV	77.7/59.3	83.7/ 61.6	81.5/57.2	84.6/ 59.8	82.9/65.3	89.0/66.1
MB	63.6/47.9	75.1/54.9	75.4/54.1	77.3/ 55.9	63.3/49.7	83.4/61.6
CM	73.3/54.7	76.2/55.6	75.6/54.1	80.1/ 57.6	70.6/54.7	86.0/63.4
BC	74.0/52.9	68.7/48.5	83.2/53.8	83.8/ 55.9	76.9/58.0	74.9/50.8
ALL	76.0/56.9	79.0/57.3	82.3/57.5	83.9/ 59.3	81.1/63.2	85.0/61.9

Table 5. Quantitative results on VOT-RGBT2019 and VOT-RGBT2020 datasets. The red bold fonts and blue fonts indicate the best and the second best performance.

	VOT-RGBT2019			VOT-RGBT2020		
Trackers	Accuracy	Robustness	EAO	Accuracy	Robustness	EAO
MANet++ [65]	50.92	53.79	27.16	-	-	-
TFNet [113]	46.17	59.36	28.78	-	-	-
MaCNet [97]	54.51	59.14	30.52	-	-	-
MANet [64]	58.23	70.1	34.63	-	-	-
mfDiMP [99]	60.19	80.36	38.79	63.8	79.3	38
ADRNet [102]	62.18	76.57	39.59	-	-	-
SiamCDA [105]	58.2	75.7	42.46	-	-	-
JMMAC [103]	66.49	82.11	48.26	66.2	81.8	42

can automatically detect and initialize the object to be tracked in subsequent frames, eliminating the need for manual annotation of the first frame. In addition, integrating detection and tracking can improve the robustness and accuracy of the tracker, as it can handle various scenarios where the object may temporarily disappear or reappear. Therefore, integrated detection and tracking is an essential direction for the development of RGBT tracking, and there is great potential for further exploration and improvement in this field.

7.3 Unsupervised and weakly supervised on RGBT tracking

Unsupervised RGBT target tracking is in its early developmental stages [59]. The exploration of unsupervised and weakly supervised methods in RGBT tracking holds significant promise for the advancement of practical and robust tracking systems. These methods offer the advantage of reducing or eliminating the need for costly and time-consuming manual annotations, particularly beneficial in real-world scenarios. Integrating unsupervised and weakly supervised learning enables tracking systems to learn from diverse, unstructured data, enhancing adaptability to dynamic environments. However, challenges persist, including designing effective feature representations, improving model interpretability,

Table 6. Challenge-based precision and success scores of 6 trackers on LasHeR (refering to Table 1 for detailed information), including MANet, MacNet, DMCNet, MANet++, SGT ,mfDiMP. The red bold fonts and blue fonts indicate the best and the second best performance.

Method	SGT [53]	MANet++ [65]	MANet [64]	MacNet [97]	mfDiMP [99]	DMCNet [66]
NO	49.9/33.1	68.3/44.9	67.4/47.2	69.2/49.2	73.8/57.2	68.1/47.5
PO	39.2/27.2	49.8/34.3	50.9/36.9	51.7/37.3	47.2/36.2	52.7/38.0
TO	30.2/21.2	38.2/26.5	41.0/29.4	42.8/30.7	34.1/26.1	42.6/30.9
HO	21.7/22.7	25.2/25.3	31.2/31.4	32.1/32.8	20.0/24.7	29.6/29.8
OV	37.1/24.6	39.2/23.5	40.6/28.1	45.3/31.9	45.3/31.6	46.9/31.7
LI	34.4/25.0	44.1/30.3	45.0/33.3	45.2/32.9	40.8/31.0	48.1/34.8
HI	42.0/27.5	56.8/37.9	58.9/41.4	59.1/41.4	54.8/40.8	61.5/42.8
AIV	31.0/21.8	44.1/29.1	45.4/34.5	44.0/32.8	35.5/29.3	47.6/37.0
LR	39.8/24.0	52.7/31.6	53.2/33.1	53.2/33.3	47.2/31.3	56.0/35.3
DEF	35.2/28.2	43.2/32.8	42.5/34.6	44.8/36.2	43.9/35.8	47.3/37.3
BC	35.9/25.9	45.6/32.2	45.4/34.2	46.3/34.4	40.3/31.3	46.7/34.9
SA	40.0/28.1	48.2/32.8	47.9/34.5	49.5/35.6	44.5/33.9	51.1/36.7
TC	36.4/25.1	45.5/30.8	46.8/33.2	47.9/34.0	43.6/32.8	49.2/34.7
MB	36.7/25.0	48.3/31.5	50.4/34.7	52.1/36.2	47.6/35.1	52.5/36.2
CM	38.1/26.3	49.3/34.1	50.8/36.9	52.6/37.9	50.1/37.8	54.1/39.0
FL	32.7/21.4	34.4/20.2	37.1/25.2	33.2/22.4	29.2/22.6	39.3/28.3
FM	36.3/25.7	47.1/33.1	48.9/36.2	49.8/36.6	49.1/38.5	51.3/37.6
SV	40.1/26.9	51.6/35.1	53.4/37.9	54.1/38.6	52.5/40.3	54.9/38.9
ARC	30.1/22.8	40.2/28.6	40.0/30.4	41.0/31.3	43.3/34.7	43.2/32.5
ALL	39.2/26.8	52.3/35.5	53.8/38.3	54.6/39.2	52.2/39.9	55.7/39.5

and addressing domain shift issues. Ongoing research efforts are crucial to overcoming these challenges and advancing the practical applications of unsupervised and weakly supervised methods in RGBT tracking.

8 CONCLUSION

In this paper, we comprehensively review existing RGBT tracking methods and classify them into the following four categories: traditional filtering method, feature decoupling based method, feature selecting based method ,and collaborative graph tracking method. In addition, we have conducted a detailed discussion on the principles of each type of method and then selected its representative methods to elaborate on. Furthermore, we summarize the existing RGBT target tracking datasets, related evaluation indicators, and analyze the performance of existing methods under the current datasets. Finally, we put forward some prospects from three perspectives: large-scale RGBT fusion dataset, integrated detection and tracking framework, and unsupervised and weakly supervised learning. Overall, this paper provides a comprehensive overview of RGBT object tracking from the perspective of fusion. We hope that readers can easily grasp the basic development of multimodal visual tracking and provide more inspiration for readers.

9 ACKNOWLEDGMENTS

This paper is partially supported by Natural Science Foundation of China under Grant No. 62006244, and the Young Elite Scientist Sponsorship Program of China Association for Science and Technology YESS20200140, and Young Elite Scientist Sponsorship Program of Beijing Association for Science and Technology BYESS2021178, and the National Natural Science Foundation of China under Grant No.91948303 and No.62302053.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. 2016. Staple: Complementary learners for real-time tracking. In *CVPR*. 1401–1409.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *ECCVW*. 850–865.

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

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. 2010. Visual object tracking using adaptive correlation filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2544–2550.

[5] Filiz Bunyak, Kannappan Palaniappan, Sumit Kumar Nath, and Guna Seetharaman. 2007. Geodesic active contour based fusion of visible and infrared video for persistent object tracking. In *WACV*. 35–35.

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer tracking. In *CVPR*. 8126–8135.

[7] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. 2017. Visual tracking by reinforced decision making. *arXiv preprint arXiv:1702.06291* 2 (2017).

[8] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. 2019. Deep meta learning for real-time target-aware visual tracking. In *ICCV*. 911–920.

[9] Ciarán O Conaire, Noel E O'Connor, Eddie Cooke, and Alan F Smeaton. 2006. Comparison of fusion methods for thermo-visual surveillance tracking. In *ICIF*. 1–7.

[10] Ciarán Ó Conaire, Noel E O'Connor, and Alan Smeaton. 2008. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. *MVA* (2008), 483–494.

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

[12] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. 2020. High-performance long-term tracking with meta-updater. In *CVPR*. 6298–6307.

[13] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR*, Vol. 1. 886–893.

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

[15] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2017. Eco: Efficient convolution operators for tracking. In *CVPR*. 6638–6646.

[16] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *BMVC*.

[17] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. 2016. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence* 39, 8 (2016), 1561–1575.

[18] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2015. Convolutional features for correlation filter based visual tracking. In *ICCVW*. 58–66.

[19] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2015. Learning spatially regularized correlation filters for visual tracking. In *ICCV*. 4310–4318.

[20] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2016. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In *CVPR*. 1430–1438.

[21] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. 2016. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*. 472–488.

[22] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. 2014. Adaptive color attributes for real-time visual tracking. In *CVPR*. 1090–1097.

[23] Heng Fan and Haibin Ling. 2019. Siamese cascaded region proposal networks for real-time visual tracking. In *CVPR*. 7952–7961.

[24] Mingzheng Feng and Jianbo Su. 2023. Learning Multi-Layer Attention Aggregation Siamese Network for Robust RGBT Tracking. *IEEE Transactions on Multimedia* (2023).

[25] Zhihong Fu, Zehua Fu, Qingjie Liu, Wenrui Cai, and Yunhong Wang. 2022. SparseTT: Visual tracking with sparse transformers. *arXiv preprint arXiv:2205.03776* (2022).

[26] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. 2019. Deep adaptive fusion network for high performance RGBT tracking. In *ICCVW*. 0–0.

[27] Chang Guo, Dedong Yang, Chang Li, and Peng Song. 2022. Dual Siamese network for RGBT tracking via fusing predicted position maps. *TVC* 38, 7 (2022), 2555–2567.

[28] Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. 2021. Graph attention tracking. In *CVPR*. 9543–9552.

[29] Jun;Cui Ying;Wang Zhenhua;Chen Shengyong Guo, Dongyan;Wang. 2020. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *CVPR*. 6269–6277.

[30] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. 2017. Learning dynamic siamese network for visual object tracking. In *ICCV*. 1763–1771.

[31] Joao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*. Springer, 702–715.

[32] Ruichao Hou, Boyue Xu, Tongwei Ren, and Gangshan Wu. 2023. MTNet: Learning Modality-aware Representation with Transformer for RGBT Tracking. In *ICME*. 1163–1168.

[33] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2020. Globaltrack: A simple and strong baseline for long-term tracking. In *AAAI*, Vol. 34. 11037–11044.

[34] Michael Isard and Andrew Blake. 1998. CONDENSATION–conditional density propagation for visual tracking. *IJCV* 29, 1 (1998), 5.

[35] Zedu Chen;Bineng Zhong;Guorong Li;Shengping Zhang;Rongrong Ji. 2020. Siamese box adaptive network for visual tracking. In *CVPR*. 6668–6677.

[36] Peng Jingchao, Zhao Haitao, Hu Zhengwei, Zhuang Yi, and Wang Bofan. 2021. Siamese infrared and visible light fusion network for RGB-T tracking. *arXiv preprint arXiv:2103.07302* (2021).

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

[38] Xiangyuan Lan, Mang Ye, Shengping Zhang, Huiyu Zhou, and Pong C Yuen. 2020. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *PRL* 130 (2020), 12–20.

[39] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*. 4282–4291.

[40] Bi Li, Wenxuan Xie, Wenjun Zeng, and Wenyu Liu. 2019. Learning to update for object tracking with recurrent meta-learner. *TIP* 28, 7 (2019), 3624–3635.

[41] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 2018. High performance visual tracking with siamese region proposal network. In *CVPR*. 8971–8980.

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

[43] Chenglong Li, Shiyi Hu, Sihan Gao, and Jin Tang. 2016. Real-time grayscale-thermal tracking via laplacian sparse representation. In *MMM*. Springer, 54–65.

[44] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2018. RGB-T Object Tracking: Benchmark and Baseline. *CoRR* abs/1805.08982 (2018).

[45] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2019. RGB-T object tracking: Benchmark and baseline. *PR* 96 (2019), 106977.

[46] Chenglong Li, Liang Lin, Wangmeng Zuo, and Jin Tang. 2017. Learning patch-based dynamic graph for visual tracking. In *AAAI*, Vol. 31.

[47] Chenglong Li, Liang Lin, Wangmeng Zuo, Jin Tang, and Ming-Hsuan Yang. 2018. Visual tracking via dynamic graph learning. *TPAMI* 41, 11 (2018), 2770–2782.

[48] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. 2020. Challenge-aware RGBT tracking. In *ECCV*. Springer, 222–237.

[49] Chenglong Li, Xiang Sun, Xiao Wang, Lei Zhang, and Jin Tang. 2017. Grayscale-thermal object tracking via multitask laplacian sparse representation. *TSMC-S* 47, 4 (2017), 673–681.

[50] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing* 281 (2018), 78–85.

[51] Chenglong Li, Zhiqiang Xiang, Jin Tang, Bin Luo, and Futian Wang. 2021. Rgbt tracking via noise-robust cross-modal ranking. *TNNLS* 33, 9 (2021), 5019–5031.

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

[53] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *ACMM*. 1856–1864.

[54] Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. 2018. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In *ECCV*. 808–823.

[55] Chenglong Li, Chengli Zhu, Jian Zhang, Bin Luo, Xiaohao Wu, and Jin Tang. 2018. Learning local-global multi-graph descriptors for RGB-T object tracking. *TCSVT* 29, 10 (2018), 2913–2926.

[56] Chenglong Li, Chengli Zhu, Shaofei Zheng, Bin Luo, and Jing Tang. 2018. Two-stage modality-graphs regularized manifold ranking for RGB-T tracking. *SPIC* 68 (2018), 207–217.

[57] Ning Wang;Wengang Zhou;Jie Wang;Houqiang Li. 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*. 1571–1580.

[58] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2019. Gradnet: Gradient-guided network for visual object tracking. In *ICCV*. 6162–6171.

[59] Shenglan Li, Rui Yao, Yong Zhou, Hancheng Zhu, Bing Liu, Jiaqi Zhao, and Zhiwen Shao. 2023. Unsupervised RGB-T object tracking with attentional multi-modal feature fusion. *Multimedia Tools and Applications* (2023), 1–19.

[60] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *CVPR*. 510–519.

[61] Yang Li, Jianke Zhu, et al. 2014. A scale adaptive kernel correlation filter tracker with feature integration.. In *ECCVW*, Vol. 8926. Citeseer, 254–265.

[62] T. Lindeberg. 2012. Scale Invariant Feature Transform. *Scholarpedia* 7, 5 (2012), 10491.

[63] Lei Liu, Chenglong Li, Yun Xiao, and Jin Tang. 2023. Quality-Aware RGBT Tracking via Supervised Reliability Learning and Weighted Residual Guidance. In *ACMM*. 3129–3137.

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

[65] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. 2021. RGBT tracking via multi-adapter network with hierarchical divergence loss. *TIP* 30 (2021), 5613–5625.

[66] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. 2022. Duality-gated mutual condition network for RGBT tracking. *TNNLS* (2022).

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

[68] Xue Mei and Haibin Ling. 2009. Robust visual tracking using l1 minimization. In *ICCV*. 1436–1443.

[69] Hyeonseob Nam and Bohyung Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*. 4293–4302.

[70] Renaud Péteri and Ondřej Šiler. 2009. Object tracking using joint visible and thermal infrared video sequences. (2009).

[71] Yuankai Qi, Shengping Zhang, Weigang Zhang, Li Su, Qingming Huang, and Ming-Hsuan Yang. 2019. Learning attribute-specific representations for visual tracking. In *AAAI*, Vol. 33. 8835–8842.

[72] Longfeng Shen, Xiaoxiao Wang, Lei Liu, Bin Hou, Yulei Jian, Jin Tang, and Bin Luo. 2022. RGBT tracking based on cooperative low-rank graph model. *Neurocomputing* 492 (2022), 370–381.

[73] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. 2013. Visual tracking: An experimental survey. *TPAMI* 36, 7 (2013), 1442–1468.

[74] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. 2017. Crest: Convolutional residual learning for visual tracking. In *ICCV*. 2555–2564.

[75] Zhangyong Tang, Tianyang Xu, and Xiao-Jun Wu. 2022. A Survey for Deep RGBT Tracking. *arXiv preprint arXiv:2201.09296* (2022).

[76] Zhengzheng Tu, Chun Lin, Wei Zhao, Chenglong Li, and Jin Tang. 2021. M 5 l: multi-modal multi-margin metric learning for RGBT tracking. *TIP* 31 (2021), 85–98.

[77] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. 2020. Siam r-cnn: Visual tracking by re-detection. In *CVPR*. 6578–6588.

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

[79] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. 2015. Visual tracking with fully convolutional networks. In *ICCV*. 3119–3127.

[80] Yulong Wang, Chenglong Li, and Jin Tang. 2018. Learning soft-consistent correlation filters for RGB-T object tracking. In *CVPR*. 295–306.

[81] Yi Wu, Erik Blasch, Genshe Chen, Li Bai, and Haibin Ling. 2011. Multiple source data fusion via sparse representation for robust visual tracking. In *ICIF*. IEEE, 1–8.

[82] Yutao Cui;Cheng Jiang;Limin Wang;Gangshan Wu. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*. 13608–13618.

[83] Gang Xiao, Xiao Yun, and JianMin Wu. 2012. A multi-cue mean-shift target tracking approach based on fuzzified region dynamic image fusion. *SCI CHINA INFORM SCI* 55 (2012), 577–589.

[84] Gang Xiao, Xiao Yun, and Jianmin Wu. 2016. A new tracking approach for visible and infrared sequences based on tracking-before-fusion. *IJDC* 4 (2016), 40–51.

[85] Xianbing Xiao, Xingzhong Xiong, Fanqin Meng, and Zhen Chen. 2023. Multi-scale feature interactive fusion network for rgbt tracking. *Sensors* 23, 7 (2023), 3410.

[86] YUN Xiao, Zhongliang Jing, Gang Xiao, JIN Bo, and Canlong Zhang. 2016. A compressive tracking based on time-space Kalman fusion model. *Information Sciences* 59, 012106 (2016), 1–012106.

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

[88] Qin Xu, Yiming Mei, Jinpei Liu, and Chenglong Li. 2021. Multimodal cross-layer bilinear pooling for RGBT tracking. *TMM* 24 (2021), 567–580.

[89] Yingjian Xue, Jianwei Zhang, Zhoujin Lin, Chenglong Li, Bihan Huo, and Yan Zhang. 2023. SiamCAF: Complementary Attention Fusion-Based Siamese Network for RGBT Tracking. *Remote Sensing* 15, 13 (2023), 3252.

[90] Tianyu Yang and Antoni B Chan. 2018. Learning dynamic memory networks for object tracking. In *ECCV*. 152–167.

[91] Zikai Song;Junqing Yu;Yi–Ping Phoebe Chen;Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *CVPR*. 8791–8800.

[92] Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. *CSUR* 38, 4 (2006), 13–es.

[93] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. 2020. Deformable siamese attention networks for visual object tracking. In *CVPR*. 6728–6737.

[94] Xiao Yun, Yanjing Sun, Xuanxuan Yang, and Nannan Lu. 2019. Discriminative fusion correlation learning for visible and infrared tracking. *MATH PROBL ENG* 2019 (2019).

[95] Sulan Zhai, Pengpeng Shao, Xinyan Liang, and Xin Wang. 2019. Fast RGB-T tracking via cross-modal correlation filters. *Neurocomputing* 334 (2019), 172–181.

[96] Fan Zhang, Hanwei Peng, Lingli Yu, Yuqian Zhao, and Baifan Chen. 2023. Dual-Modality Space-Time Memory Network for RGBT Tracking. *IEEE Transactions on Instrumentation and Measurement* (2023).

[97] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. 2020. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors* 20, 2 (2020), 393.

[98] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang. 2014. Fast compressive tracking. *IEEE transactions on pattern analysis and machine intelligence* 36, 10 (2014), 2002–2015.

[99] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. 2019. Multi-modal fusion for end-to-end rgb-t tracking. In *ICCVW*. 0–0.

[100] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. 2019. Learning the model update for siamese trackers. In *ICCV*. 4010–4019.

[101] Pengyu Zhang, Dong Wang, and Huchuan Lu. 2020. Multi-modal visual tracking: Review and experimental comparison. *arXiv preprint arXiv:2012.04176* (2020).

[102] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 2021. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *IJCV* 129 (2021), 2714–2729.

[103] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 2021. Jointly modeling motion and appearance cues for robust RGB-T tracking. *TIP* 30 (2021), 3335–3347.

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

[105] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. 2021. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network. *TCSVT* 32, 3 (2021), 1403–1417.

[106] Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. 2020. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion* 63 (2020), 166–187.

[107] Xingchen Zhang, Ping Ye, Shengyun Peng, Jun Liu, Ke Gong, and Gang Xiao. 2019. SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks. *IEEE Access* 7 (2019), 122122–122133.

[108] Xingchen Zhang, Ping Ye, Shengyun Peng, Jun Liu, and Gang Xiao. 2020. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *SPIC* 84 (2020), 115756.

[109] Jie Zhao, Kenan Dai, Pengyu Zhang, Dong Wang, and Huchuan Lu. 2022. Robust Online Tracking With Meta-Updater. *TPAMI* (2022).

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

[111] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. 2019. Dense feature aggregation and pruning for RGBT tracking. In *ACMM*. 465–472.

[112] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. 2020. Quality-aware feature aggregation network for robust RGBT tracking. *TIV* 6, 1 (2020), 121–130.

[113] Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo, and Liang Wang. 2021. RGBT tracking by trident fusion network. *TCSVT* 32, 2 (2021), 579–592.

[114] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. 2018. Distractor-aware siamese networks for visual object tracking. In *ECCV*. 101–117.