Learning Multi-Layer Attention Aggregation Siamese Network for Robust RGBT Tracking

Mingzheng Feng and Jianbo Su, *Senior Member, IEEE*

***Abstract*—Recent years have witnessed the popularity of integrating Siamese network into RGBT tracking for fast-tracking. However, these trackers mostly utilize the feature information of the last output layer and ignore the benefits of multi-layer information. In addition, they often adopt feature-level fusion for different modalities but fail to explore the strength of decision-level fusion, which may easily decrease their flexibility and independence. In this article, a novel multi-layer attention aggregation Siamese network on the decision level is proposed for robust RGBT tracking. To be specific, a hierarchical channel attention Siamese network is built to recalibrate the extracted multi-layer features from RGB and thermal infrared images. This can focus on more discriminative features to learn robust feature representation. Then, a depth-wise correlation operation is performed to produce RGB and thermal response maps, respectively. To better exploit and utilize the complementary RGB and thermal information, a contribution-aware aggregation network is designed to adaptively aggregate them. Lastly, a classification and regression network is adopted to complete the bounding box prediction. Extensive experiments on four large-scale RGBT benchmarks demonstrate outstanding tracking ability over other state-of-the-art trackers.**

***Index Terms*—RGBT tracking, hierarchical attention network, contribution-aware aggregation network.**

I. Introduction

VISUAL tracking [1], [2], [3] has received widespread attention for its considerable potential in numerous practical applications, such as intelligent surveillance and self-driving systems. The purpose of visual tracking is to estimate the position and shape of the target in sequential frames based on its initial state. Recently, trackers based on Siamese networks are extensively designed due to their strong balance between accuracy and speed. As pioneering work, Bertinetto et al. [4] design fully convolutional Siamese network to transform the tracking task into paradigm matching for highly efficient tracking. Encouraged by its success, more and more trackers have been proposed to further exploit the potentiality of the Siamese network by building different Siamese architectures [5], learning efficient Siamese network [6], [7], introducing deep reinforcement learning [8], [9], utilizing powerful training loss [10], and so on [11], [12], [13], [14]. For instance, Han et al. [15] propose asymmetric convolution in Siamese-based trackers to learn powerful feature matching for robust tracking. Dong et al. [8], [9] improve the current deep reinforcement learning method to better learn the hyperparameters of the tracker, thereby boosting the tracking performance. Although these trackers based on RGB information have made rapid progress in visual tracking, their performance [16], [17] needs to be further improved facing complex scenarios such as illumination change and haze interference due to the limitation of the inherent defects of RGB images. Recently, RGBT tracking [18], [19] combining complementary RGB and thermal infrared information has aroused extensive research interest. On one hand, thermal infrared images are insensitive to illumination conditions [20], [21] and can capture the target at night and foggy days. On the other hand, RGB images can show richer color and texture information for foreground-background separation when facing thermal crossover. Hence, RGBT tracking can better guarantee the tracking robustness encountering complex challenging scenes.

Existing RGBT trackers can be divided into two aspects according to the feature type. One is to utilize manually extracted features. Li et al. [22] build an adaptive collaborative sparse model to incorporate extracted features from grayscale and thermal videos. Feng et al. [23] design a differentiated extraction scheme and an adaptive weighting scheme to obtain and fuse the features from different modalities. However, their performance is decreased facing some challenging scenarios, such as background clutter and occlusion for that they cannot provide more discriminating feature information. Inspired by the impressive performance of convolution neural network (CNN) in computer vision fields, there are many studies using CNN features to improve tracking performance. Li et al. [24] build a two-stream convolution network to obtain rich feature information and then build a FusionNet to adaptively fuse them. Zhu et al. [25] design a fully convolutional dense information aggregation module to obtain and integrate the information from RGB and thermal modalities. These RGBT trackers based on the CNN framework generally have a distinct advantage over traditional ones in that they can learn robust feature representation. However, their speed is nowhere near that of real-time application.

Recent studies have focused on the incorporation of Siamese network into RGBT tracking to improve tracking efficiency. On the one hand, the structure of Siamese network itself is simple.

Manuscript received 9 February 2023; revised 12 June 2023; accepted 18 August 2023. Date of publication 30 August 2023; date of current version 14 February 2024. This work was supported in part by the key Project of the National Natural Science Foundation of China under Grant 91748120, and in part by Shanghai Cross-disciplinary Research Fund under Grant JYJC202214. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Jianlong Fu. *(Corresponding author: Jianbo Su.)*

The authors are with the Key Laboratory of System Control and Information Processing, Department of Automation, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: fmz@sjtu.edu.cn; jbsu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3310295

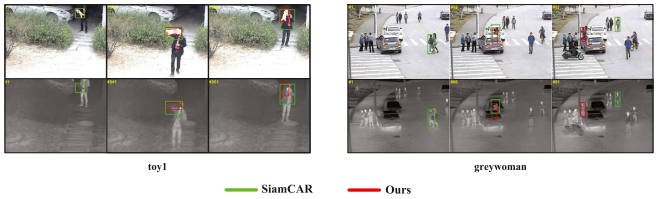

Fig. 1. Qualitative results of our multi-layer attention aggregation network incorporated into the baseline tracker SiamCAR [30]. By adjusting the baseline tracker SiamCAR in the first frame, we significantly enhance its tracking ability in challenging scenarios such as huge deformation (left) or severe occlusion (right).

The input template and the search images are transformed by the feature extractor with shared parameters, which can learn useful similar matching information more effectively to improve the robustness of tracking. On the other hand, Siamese network learns the similarity between the template and the search areas through end-to-end offline training, which overcomes the limitations of pre-trained CNN and can estimate the similarity online in real-time during the tracking process. Zhang et al. [26] introduce convolutional Siamese network into RGBT tracking, achieving a running speed of about 30 FPS. This is also the first work to complete RGBT tracking by Siamese network. Furthermore, Zhang et al. [27] apply the dynamic Siamese network to generate and fuse cross-layer features for robust RGBT tracking. Although they are faster than other advanced RGBT trackers, their tracking performance is much lower.

In this article, a multi-layer attention aggregation Siamese network is designed for RGBT tracking, which can achieve superior performance and maintain fast running speed. To obtain more discriminative semantic information at different levels, a modified dual-stream ResNet-50 network is adopted to extract hierarchical target features from RGB and thermal images. For each modal branch, a channel attention network is utilized to explicitly model the interdependencies between feature channels, which can emphasize the powerful features and improve the robustness of features. To make the generated response map retain more useful information for subsequent tasks, a depth-wise cross-correlation layer is introduced to calculate the cross-correlation between feature maps channel by channel. In order to fully utilize the complementary advantages of different modalities, a contribution-aware aggregation network is designed to adaptive learn the reliable weight of RGB and thermal modalities for fusion. Finally, the fused response is passed through the classification and regression network to locate the target. Evaluation results on several popular RGBT tracking benchmarks show that the designed multi-layer attention aggregation Siamese network can significantly improve tracking performance. To better prove the effectiveness of the designed multi-layer attention aggregation Siamese network, we compare our proposed tracker with the base model SiamCAR [28] in Fig. 1. From Fig. 1, we can see that our proposed tracker can accurately track the target facing challenging scenarios such as large appearance change or severe occlusion.

The main contributions of this work are summarized as:

- A novel multi-layer attention aggregation Siamese network is proposed for RGBT tracking, which can achieve high performance and maintain fast running speed.
- A hierarchical interaction channel attention network is designed to provide multilevel abstractions of target and recalibrate the feature channels of the multi-layer features, which can learn more robust feature representation.
- A contribution-aware aggregation network is designed to adaptive fuse the responses of RGB and thermal modalities to take full use of their complementary advantages.
- Extensive experiments on four challenging benchmarks, GTOT [22], RGBT210 [29], RGBT234 [30], and LasHeR [31] demonstrate the superior performance against other advanced trackers.

II. Related Work

This section briefly reviews some of the most relevant available results for our work: RGB tracking, RGBT tracking, and Multimodal fusion.

A. RGB Tracking Methods

Benefiting from the popularity of CNN in the computer vision field, many RGB trackers based on deep convolutional feature have been put forward. Danelljan et al. [32] train a two-layer convolutional network to classify the tracking target and maintain high tracking robustness in the face of distractors. Bhat et al. [33] propose an end-to-end training network to improve the discriminative ability of the predicted model for robust tracking. In general, they can achieve excellent tracking results at relatively slow speed. Differently, Siamese-based trackers have drawn prevalent attention for that they can better balance accuracy and efficiency. SiamFC [4] trains the first fully convolutional Siamese network to estimate the similarity between the template and search region, achieving outstanding tracking results. DSiam [34] builds a dynamic Siamese network to deal with the target appearance variation and suppress background interference. To obtain more discriminative features, SiamFC-tri [10] proposes a triplet loss in Siamese network to replace the pairwise loss for training. Quadruplet [5] designs a quadruplet deep network to sufficiently exploit the potential connection of training samples. CFNet-Hy [9] introduces a deep Q-learning strategy to adaptively optimize model hyperparameters. However, many Siamese-based trackers are not good at solving the case of sharp changes in the aspect ratio of the target. SiamRPN [35] proposes Siamese region proposal network for proposal feature extraction to more accurately predict the target boundary box. CLNet [7] designs compact latent network to fully learn sequence-specific information, aiming to better cope with changing scenarios. Recently, attention mechanism has been introduced in visual tasks [36], [37] to further obtain powerful representation. For example, HASiam [11] introduces the attention mechanism into the Siamese network to enhance its matching discrimination for robust tracking. LSSiam [12] proposes a local semantic Siamese network to learn more robust features for coping with model drift and designs a generally focal logistic loss to further obtain

strong feature representation. To further boost the tracking performance, DaSiamRPN [38] develops distractor aware training to increase the discrimination of the tracker. SiamRPN++ [39] and SiamDW [40] remove the influence factors such as padding and adopt modern deep networks such as ResNet [41] as backbone to improve the tracking accuracy. However, they often require heuristic configuration and rich empirical tricks to achieve desired performance. Aims to this problem, SiamBAN [2] and SiamCAR [28] utilize the expressive capability of fully convolutional network to abandon tricky hyper-parameter tuning of anchors and make the tracking process more simple.

B. RGBT Tracking Methods

RGBT tracking has attracted widespread research interest in recent years for its all-weather robust tracking performance. Early RGBT trackers mostly pay attention to sparse representation with the ability of noise suppression. Liu et al. [42] apply joint sparse representation to obtain the similarity of different modalities and fuse them by minimizing the sparse representation coefficient. Li et al. [29] design a weighted regularized graph model to obtain more discriminative features. In addition, RGBT tracking based on correlation filter has emerged for high efficiency. Zhang et al. [43] utilize the low-rank constraint to collaboratively learn filters. Luo et al. [44] design a hybrid tracking-by-detection architecture to exploit the complementary benefits of different modalities. But these trackers only focus on hand-crafted features, which cannot handle complex challenging scenarios. Recently, RGBT tracking based on deep learning presents competitive performance for the strong feature representation brought by CNN. MANet [45] designs a trained deep network composed of three types of adapters to generate discriminative deep representations. mfDiMP [46] trains an end-to-end network by discriminative loss and explores the fusion schemes in different levels to obtain the best fusion framework. These trackers achieve high tracking performance, but their tracking efficiency needs to be further improved. Currently, RGBT tracker based on Siamese network has gradually become a research hotspot for its end-to-end training capability and high efficiency. Guo et al. [47] design a dual Siamese subnetwork and a joint modal attention mechanism to effectively integrate the information between different modalities. Zhang et al. [27] introduce a dynamic Siamese network to obtain more robust feature information. Although they have excellent tracking speed, there is still a large gap in accuracy against other advanced RGBT trackers.

C. Multimodal Fusion

The aim of multimodal fusion is to fuse information from different modalities to obtain more reliable output. In terms of the fusion stage, multimodal fusion can be divided into three types, including pix-level, feature-level, and decision-level fusion. In general, pix-level fusion is fulfilled by simply concatenating. Cvejic et al. [48] explore different methods to fuse RGB and infrared information on pixel-level. Wu et al. [49] simply concatenate the image blocks of different modalities into a one-dimensional vector. Pixel-level fusion is relatively simple but easily disturbed by noise information. Feature-level fusion is the most popular type of multimodal fusion. Zhu et al. [50] design an adaptive fusion scheme to integrate deep feature information of different modalities, which can handle significant appearance changes. Wang et al. [51] introduce a transformation scheme to learn the reliable features between different modalities and then fuse them through element-by-element summation. Feature-level fusion can better exploit the complementary benefits of different modalities, but the independence of information is easily decreased. As the highest fusion stage, decision-level fusion combines the responses of each modality to achieve adaptive fusion, which ensures the independence and flexibility of the model. Zheng et al. [52] propose a decision-level fusion scheme to query the validity of feature recognition in an adaptive manner. Jain et al. [53] design a score normalization scheme for decision fusion. However, there are seldom studies to exploit decision-level fusion in RGBT tracking. In view of it, we pay attention to decision-level fusion in this article to boost the tracking performance.

III. Multi-Layer Attention Aggregation Siamese Network

A. The Architecture Overview

As shown in Fig. 2, the designed architecture contains four main parts: Hierarchical channel attention network for feature extraction, Depth-wise cross correlation for unimodal response generation, Contribution-aware aggregation network for multimodal response fusion, classification and regression network for target prediction.

Firstly, RGB and thermal images are fed into the two-stream Siamese network to complete multi-layer feature extraction. To learn stronger discriminative features, a channel attention network is introduced to recalibrate the feature channel by explicitly modeling the interdependencies. Then, the recalibrated features are input into a depth-wise correlation module for generating RGB and thermal responses. By replacing ordinary cross-correlation with depth-wise cross-correlation, we can obtain multiple semantic similarity maps that retain rich information, so as to determine the target location and scale more accurately in the subsequent prediction subnetworks. Then, to better exploit and utilize the complementary advantage of RGB and thermal information, a contribution-aware aggregation network is designed to adaptively learn the weights of RGB and thermal modalities for response fusion. Finally, the fused response map passes through the classification and regression network for target prediction. In the following, we describe each part in detail, including the hierarchical channel attention network, depth-wise cross-correlation module, contribution-aware aggregation network, and classification and regression network.

B. Hierarchical Channel Attention Network for Feature Extraction

As shown in Fig. 3, the hierarchical channel attention network consists of a backbone network and a channel attention network (CAN). Specifically, the backbone network consists of

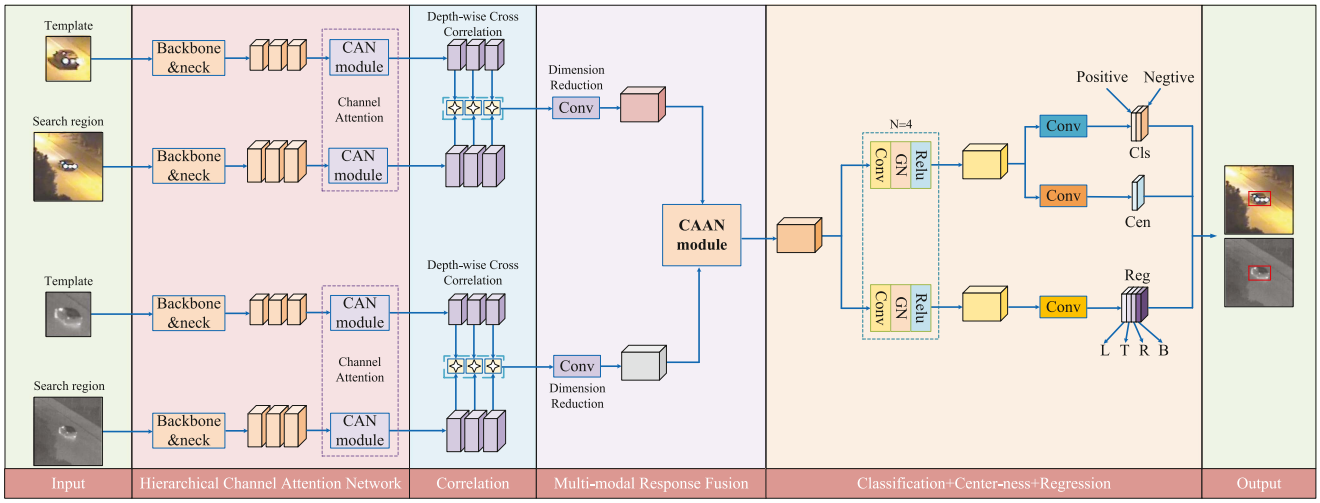

Fig. 2. Overall network architecture of the designed RGBT tracker. It mainly contains four parts: Hierarchical channel attention network for feature extraction, depth-wise cross correlation for unimodal response generation, contribution-aware agregation network for multi-modal response fusion, classification, and regression subnetwork for target prediction.

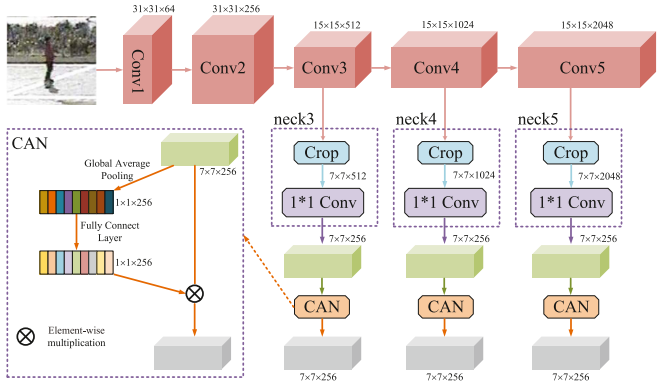

Fig. 3. Illustration of hierarchical channel attention network in RGB template branch. It is composed of a backbone&neck network and a channel attention network (CAN).

ResNet-50 [41]. To improve the feature resolution, we remove the downsampling operations in the conv4 and conv5 blocks and set their stride to 1. At the same time, their atrous rate are set to 2 and 4 respectively to enlarge the receiving field. Moreover, the deep features in low-level and high-level represent different attribute information of target, which are useful for discrimination. There, we select the features in the conv3, conv4, and conv5 blocks of the backbone to boost the tracking performance. To effectively reduce the amount of computation, a 1×1 convolution operation is added at the end of them. In addition, each Siamese network is composed of template and search branch. For convenience, the input in template and search branch are denoted as x and z. The output features of the backbone network are defined as $\varphi_{RGB}(x_{RGB})$, $\varphi_T(x_T)$, $\varphi_{RGB}(z_{RGB})$ and $\varphi_T(z_T)$, respectively.

A central theme of visual tracking is the search for more powerful representations that capture only those properties of an image that are most salient for a given target, enabling improved tracking performance. Inspired by [54], a channel attention network is introduced to hierarchical interaction Siamese network, which can emphasize the more discriminative features and improve the robustness of features. As seen in the lower left part of Fig. 3, the channel attention network uses Global Average Pooling (GAP) and fully connected layers to obtain the weight of each channel and then complete the feature update. The GAP can improve the localization ability of convolutional neural network and FC can better fit the complex correlation between feature channels. We first utilize GAP to get each compressed feature channel m_c as:

$$m_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \varphi_c(i,j), \tag{1}$$

where H and W represent the height and width of a feature map, and φ_c is the c-th channel feature map. Then the weight coefficient w_c representing the reliability of channel features is obtained as:

$$w_c = \sigma\left(fc_2\left(\delta fc_1\left(m_c\right)\right)\right), \tag{2}$$

where fc_1 and fc_2 denote two different fully connected layers, σ and δ are sigmoid and ReLU activation function, respectively. Finally, the weight coefficient w_c and the original features φ_c are multiplied to obtain the recalibrate feature:

$$\varphi_c' = w_c \cdot \varphi_c, \tag{3}$$

where φ_c' is the weighted c-th channel feature map. The channel attention network is utilized on each feature layer to pay attention to the significant regions which can boost the tracking performance.

C. Depth-Wise Cross-Correlation for Unimodal Response Generation

After obtaining the feature information of template and search branches, most of the Siamese-based trackers adopt cross-correlation operation to produce a single-channel response map

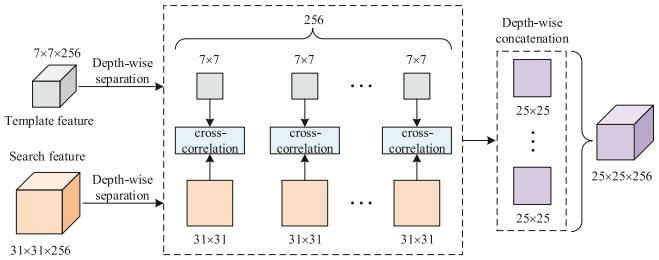

Fig. 4. Illustration of depth-wise correlation layer. It is composed of depth-wise separation, cross-correlation, and depth-wise concatenation.

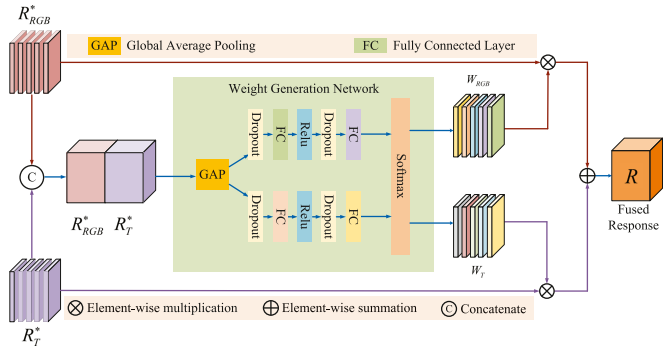

Fig. 5. Overall architecture of the contribution-aware aggregation network. The contribution-aware aggregation network is mainly composed of global average pooling layer, fully connected layer, Relu and Softmax activation function.

which easily leaves out important information for that different feature channels contain diverse semantic information. Therefore, we hope that the generated response map retains as much useful information as possible. Inspired by [39], a depth-wise correlation layer is introduced in our work to calculate the cross-correlation between feature maps channel by channel. As shown in Fig. 4, the depth-wise correlation layer consists of three parts: depth-wise separation, cross-correlation, and depth-wise concatenation. To be specific, the template feature and search feature are firstly separated into 256 two-dimensional matrices in depth-wise according to the number of feature channels. Then each two-dimensional matrice of the template feature and search feature is cross-correlated to obtain the corresponding response results. Finally, the corresponding response results are concatenated in depth-wise to produce multiple-channel response maps. By replacing ordinary cross-correlation with depth-wise cross-correlation, we can obtain multiple semantic similarity maps that retain rich information, so as to determine the target location and scale more accurately in the subsequent prediction subnetworks. The specific formulas are expressed as follows:

$$R_{RGB} = \varphi'_{RGB}\left(z_{RGB}\right) \propto \varphi'_{RGB}\left(x_{RGB}\right), \tag{4}$$

$$R_T = \varphi'_T\left(z_T\right) \propto \varphi'_T\left(x_T\right), \tag{5}$$

where $\propto$ defines the channel-by-channel correlation operation, as shown in Fig. 4. $\varphi'_{RGB}(x_{RGB})$, $\varphi'_{RGB}(z_{RGB})$, $\varphi'_T(x_T)$ and $\varphi'_T(z_T)$ are the recalibrated features by the channel attention network. R_{RGB} and R_T represents the response map of RGB and thermal modalities. Then, three outputs response denoted respectively as $R_{RGB,i=3:5}$ and $R_{T,i=3:5}$ are concatenated as a unity:

$$R^*_{RGB} = \mathrm{cat}\left(R_{RGB,3}, R_{RGB,4}R_{RGB,5}\right), \tag{6}$$

$$R^*_T = \mathrm{cat}\left(R_{T,3}, R_{T,4}R_{T,5}\right), \tag{7}$$

where R^*_{RGB} and R^*_T contains 3×256 channels, respectively. To effectively reduce the amount of computation, the response map R^*_{RGB} and R^*_T are convoluted with a 1×1 kernel to change the channel dimension to 256, respectively. Finally, R^*_{RGB} and R^*_T are adopted as the input to the contribution-aware aggregation network for multi-modal response fusion.

D. Contribution-Aware Aggregation Network for Response Fusion

After obtaining R^*_{RGB} and R^*_T from RGB and thermal branches, we need to obtain the final response map for subsequent tracking tasks. Therefore, how to fuse them to better utilize the complementary advantages of different modalities is a particularly important problem for RGBT tracking. Most existing fusion strategies simply concatenate them or sum them in element-wise. However, they tend to ignore the contributions of different modalities in certain video sequences. To effectively fuse different modal information and boost tracking performance, a contribution-aware aggregation network is presented in Fig. 5, which can adaptively learn the reliability of each modality.

In the contribution-aware aggregation network, R^*_{RGB} and R^*_T are first concatenated and then sent to the weight generation network to generate the weights of each modality. The weight generation network consists of an average pooling layer, two independent and parallel full connection layers, Relu and Softmax activation functions. The specific calculation process is shown as:

$$w_{RGB} = \zeta\left(fc^{rgb2}\left(\delta\left(fc^{rgb1}\left(GAP\left(\mathrm{cat}\left(R^*_{RGB}, R^*_T\right)\right)\right)\right)\right)\right), \tag{8}$$

$$w_T = \zeta\left(fc^{t2}\left(\delta\left(fc^{t1}\left(GAP\left(\mathrm{cat}\left(R^*_{RGB}, R^*_T\right)\right)\right)\right)\right)\right), \tag{9}$$

where $\mathrm{cat}(\cdot)$ represents the concatenating operation; $GAP(\cdot)$ represents the global average pooling function; $fc(\cdot)$ represents fully connected module including dropout, linearly connected; δ and ζ is the ReLU activation function and softmax activation function, respectively; w_{RGB} and w_T are the contribution weights of RGB and thermal modalities.

Finally, the fused response map R can be calculated as follows:

$$R = w_{RGB} \cdot R^*_{RGB} + w_T \cdot R^*_T \tag{10}$$

E. Classification and Regression Network for Target Prediction

The structure of the classification and regression network is shown in Fig. 2. It consists of three branches: a classification branch to determine the category of each position, a center-ness branch to suppress low-quality bounding boxes, and a regression branch to determine target bounding.

For the classification branch, the binary cross entropy (BCE) loss is adopted to classify each domain. The specific calculation process of BCE loss is as follows:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1-y_i)\log(1-p_i), \tag{11}$$

where N denotes sample number and p_i represents the predicted probability value of the i-th sample. y_i denotes the actual sample value, where positive samples are 1 and others are 0.

For the regression branch, the Intersection over Union (IOU) loss in UnitBox [55] is adopted to calculate the regression loss as follows:

$$L_{\text{reg}} = 1 - IOU, \tag{12}$$

where IOU is the ratio of intersection to union between the predicted box and ground-truth.

For the center-ness branch, it is noted that the positions far away from the target center easily generate low-quality predicted boxes. Following [56], a center-ness branch is introduced to filter out low-quality predicted boxes and the center-ness target can be calculated as:

$$\text{center-ness} = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}, \tag{13}$$

where l^*, t^*, r^* and b^* denote the regression bounding box coordinates of the location and the range of center-ness is 0 to 1. The center-ness loss L_{cen} is trained by BCE loss.

Finally, the overall loss is calculated as:

$$L = L_{cls} + \lambda_1 L_{cen} + \lambda_2 L_{reg}, \tag{14}$$

where L_{cls}, L_{reg}, and L_{cen} denote classification, regression, and center-ness loss, respectively. λ_1 and λ_2 are the weighted constant of regression and center-ness loss, respectively.

IV. Experiments

The performance of our designed tracker and other advanced trackers is evaluated on four challenging tracking benchmarks to demonstrate its effectiveness.

A. Evaluation Settings

Datasets: GTOT is a popular RGBT dataset and contains 50 RGBT video sequences annotated with seven challenging attributes. RGBT210 is a large-scale RGBT dataset with 210 highly-aligned RGBT video sequences. It takes many challenges into consideration and is annotated with 12 attributes. RGBT234 contains 234 RGBT sequences with the longest sequence up to 4,000 frames. As with RGBT210, it is also annotated with 12 challenging attributes to facilitate the challenge-based evaluation. LasHeR is the largest RGBT dataset and contains 1224 high-aligned video sequences. 245 sequences of it are divided separately into testing datasets, and the remaining are designed as training datasets. In addition, it has richer attribute annotations such as aspect ratio change and similar appearance.

Evaluation Metric: The metrics of precision rate (PR) and success rate (SR) are utilized for performance evaluation. PR measures the percentage of frames whose output position is within a given threshold distance, and we set the threshold to 5 pixels for GTOT and 20 pixels for other datasets to compute the representative PR. SR measures the percentage of successfully tracked frames whose overlap score is higher than the predefined threshold and the representative SR is computed by the area under the curves.

Implementation Details: To facilitate comparison, we resize the input size of the template and search region to 127×127 and 255×255, respectively. The initialized parameters of the backbone are obtained by the pre-trained ResNet-50 in the Imagenet dataset [57]. During the training phase, we utilize the stochastic gradient descent (SGD) with an initial learning rate of 0.001 to optimize the designed network. The values of weight decay and momentum are set to 0.0001 and 0.9, respectively. Moreover, we train the proposed network using the LasHeR training dataset and choose the values of batch size and training iteration as 48 and 20, respectively. In calculating training loss, the weight constants λ_1 and λ_2 are set to 1 and 3, respectively. In the tracking phase, we only utilize the target state of the initial frame to identify the template patch and then sent it into the template branch. As a consequence, we can pre-compute and fix the template branch during the whole tracking process. In addition, the search area within the current frame is used as the input of the search branch, and the sample of candidate area with the highest score is used to determine the final result. Our designed tracker is conducted in Python using PyTorch on a server with Intel Xeon(R) E5-2620 CPU, 48 G RAM, Nvidia GTX 1080Ti.

B. Evaluation on GTOT Dataset

Fig. 6 shows the comparative result of our designed tracker with 12 advanced trackers including JMMAC [58], MANet [45], FANet [50], MaCNet [59], DAPNet [25], RT-MDNet+RGBT [60], DAT+RGBT [61], CMR [62], MDNet+RGBT [63], ECO [64], CCOT [65], SiamDW+RGBT [40], on GTOT dataset. We can see that our designed tracker performs best in both metrics. Specifically, the performance of our designed tracker is up to 91.3%/75.1% in PR/SR, which has 1.1%/1.9% improvement over the second-best tracker JMMAC. In addition, compared with other competitive RGBT trackers, i.e., MANet, FANet, and MaCNet, our designed tracker achieves 1.9%/2.7%, 2.2%/2.3%, and 2.7%/4.9% improvement, respectively. These results prove the designed multi-layer attention aggregation Siamese network is effective for RGBT tracking. Moreover, our designed tracker is significantly superior to advanced RGB trackers like CCOT and ECO, proving the effectiveness of integrating thermal information into RGB tracking.

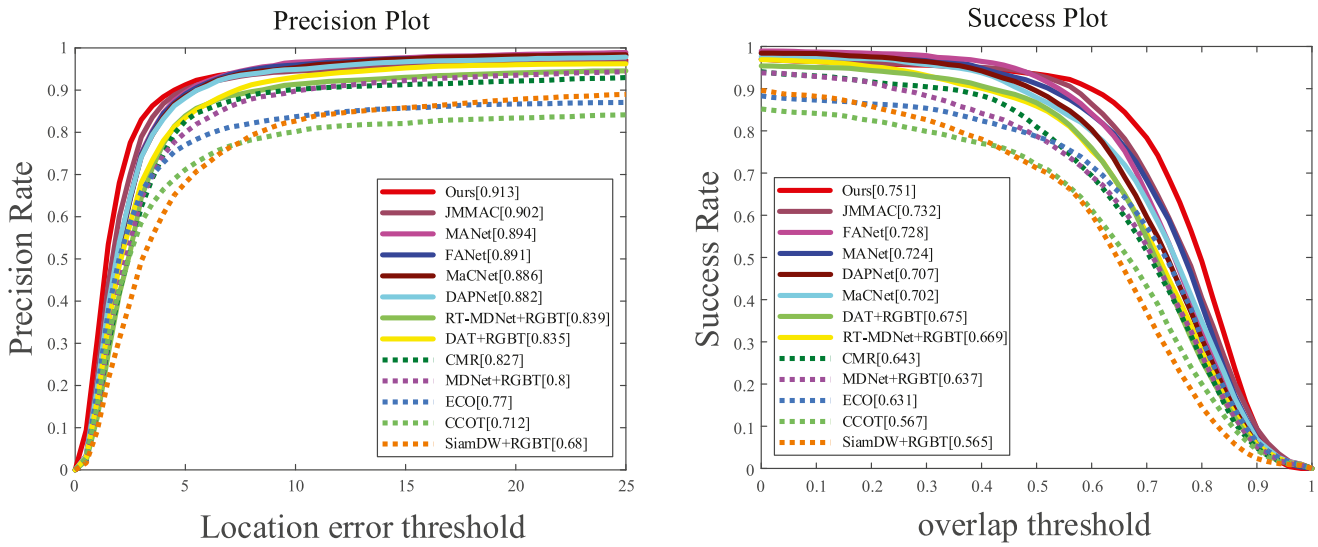

Fig. 6. Evaluation curves of different trackers on GTOT dataset.

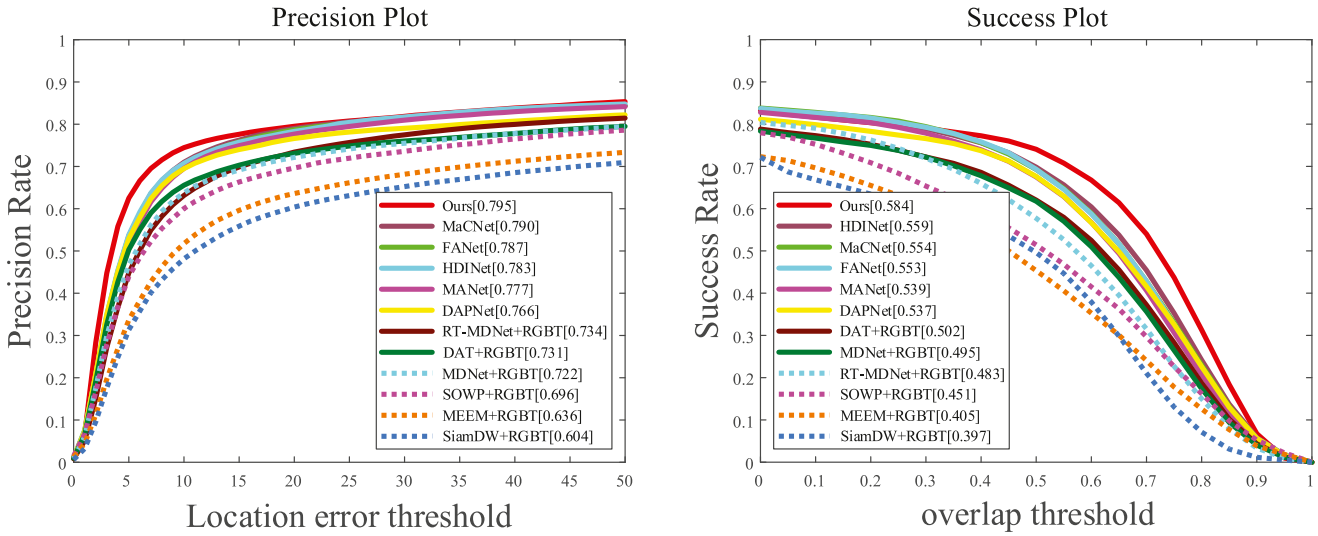

Fig. 7. Evaluation curves of different trackers on RGBT234 dataset.

C. Evaluation on RGBT234 Dataset

To demonstrate the feasibility of our designed tracker, a series of experiments are performed on RGBT234 dataset.

Overall performance: Fig. 7 shows the overall evaluation results of our designed tracker and other 11 advanced RGBT trackers including HDINet [66], MANet [45], FANet [50], MaCNet [59], DAPNet [25], RT-MDNet+RGBT [60], SOWP+RGBT [67], MDNet+RGBT [63], DAT+RGBT [61], MEEM+RGBT [68], and SiamDW+RGBT [40]. It is not hard to see that our designed tracker obtains 79.5%/58.4% in PR/SR and achieves the best performance. In particular, our designed tracker achieves 0.5% performance improvement over MaCNet ranking second in PR, and 2.5% performance improvement over HDINet ranking second in SR. While our designed tracker has less performance improvement in PR over MaCNet, it is about 27 times faster than MaCNet in speed. In addition, compared with the recent tracker FANet, our designed tracker still obtains 0.8%/3.1% performance improvement in PR/SR. These experiments validate the effectiveness of our designed multi-layer attention aggregation Siamese network for RGBT tracking.

Attribute-based performance: There are 12 challenge attribute labels on RGBT234 dataset including camera moving (CM), no occlusion (NO), low resolution (LR), thermal crossover (TC), background clutter (BC), fast motion (FM), motion blur (MB), deformation (DEF), partial occlusion (PO), heavy occlusion (HO), scale variation (SV) and low illumination (LI). Table I presents the attribute-based performance of our designed tracker and eight advanced trackers, including HDINet [66], SGT [29], MANet [45], FANet [50], MaCNet [59], DAPNet [25], MDNet+RGBT [63], and DAT+RGBT [61]. We can see that our designed tracker has obvious advantages over the other trackers in most challenges, such as CM, MB, and TC. Especially

TABLE I
ATTRIBUTE-BASED PRECISION RATE AND SUCCESS RATE ON RGBT234 DATASET OF OUR DESIGNED TRACKER AGAINST OTHER EIGHT RGBT TRACKERS

	SGT [29]	MDNet+RGBT [63]	DAT+RGBT [61]	DAPNet [25]	MANet [45]	HDINet [66]	FANet [50]	MaCNet [59]	SiamMLAA(Ours)
NO	87.7/55.5	86.2/61.1	85.9/61.3	90.0/64.4	88.7/64.6	88.4/65.1	88.2/65.7	92.7/66.5	94.7/71.3
PO	77.9/51.3	76.1/51.8	75.9/52.0	82.1/57.4	81.6/56.6	84.9/60.4	86.6/60.3	81.1/57.2	84.5/62.7
HO	59.2/39.4	61.9/42.1	64.6/43.6	66.0/45.7	68.9/46.5	67.1/47.3	66.5/45.8	70.9/48.8	67.9/48.4
LI	70.5/46.2	67.0/45.5	65.0/42.5	77.5/53.0	76.5/51.3	77.7/53.2	80.3/54.8	77.7/52.7	80.5/58.2
LR	75.1/47.6	75.9/51.5	69.7/47.3	75.0/51.0	75.7/51.5	80.1/54.5	79.5/53.2	78.3/52.3	80.8/55.3
TC	76.0/47.0	75.6/51.7	71.5/49.4	76.8/54.3	75.4/54.3	77.2/57.5	76.6/55.1	77.0/56.3	81.4/58.8
DEF	68.5/47.4	66.8/47.3	66.8/46.7	71.7/51.8	72.0/52.4	76.2/56.5	72.2/52.7	73.1/51.4	76.3/57.9
FM	67.7/40.2	58.6/36.6	63.4/38.5	67.0/44.3	69.4/44.9	71.7/47.5	68.1/43.8	72.8/47.1	74.7/54.5
SV	69.2/43.4	73.5/50.5	76.8/53.2	78.0/54.2	77.7/54.2	77.5/55.8	78.5/56.3	78.7/56.1	80.2/60.9
MB	64.7/43.6	65.4/46.3	66.8/47.5	65.3/46.7	72.6/51.6	70.8/52.6	70.0/50.3	71.6/52.5	74.9/55.9
CM	66.7/45.2	64.0/45.4	64.7/45.2	66.8/47.4	71.9/50.8	69.7/51.4	72.4/52.3	71.7/51.7	77.7/57.5
BC	65.8/41.8	64.4/43.2	69.0/45.7	71.7/48.4	73.9/48.6	71.1/47.8	75.7/50.3	77.8/50.1	75.2/51.6
ALL	72.0/47.2	72.2/49.5	73.1/50.2	76.6/53.7	77.7/53.9	78.3/55.9	78.7/55.3	79.0/55.4	79.5/58.4

The best, second, and third results are in red, green, and blue colors, respectively.

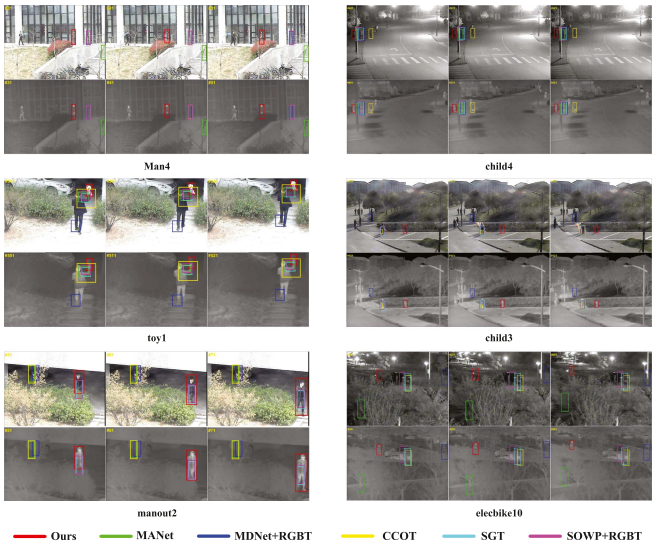

Fig. 8. Visual results of our designed tracker against other advanced trackers on six video sequences.

in the challenge of CM, the performance of the proposed tracker achieves 77.7%/57.5% and obtains 5.3%/5.2% improvement against the second-best tracker FANet. In addition, facing the challenge of TC, our designed tracker performs best and achieves 4.2%/1.3% performance improvement over the second-best tracker HDINet, which indicates that our designed tracker can effectively handle scenes with poor single-mode image quality. Furthermore, the superior performance of the designed tracker in the challenges of DEF, SV, and CM also indicates that our designed model can keep robustness when the appearance of the target changes significantly. Generally speaking, these results prove the excellent tracking capability of our designed tracker in dealing with various challenges.

Visual comparison: Fig. 8 shows the visual tracking results of our designed tracker and other five advanced RGBT trackers, including MANet [45], MDNet+RGBT [63], CCOT [65], SGT [29] and SOWP+RGBT [67], on six sequences. It is intuitive to see that our designed tracker has better positioning ability in complex challenges, such as background clutter, thermal crossover, and low illumination. For example, in the sequence *Man4*, our designed tracker is able to accurately predict the target and perform better in the challenges of background clutter and occlusion. In the sequence *child4*, *elecbike10*, and *manout2*, when the target faces the interference of low illumination and low resolution, our designed tracker can successfully track the target while others lose it. Overall, the visual comparison results demonstrate our designed tracker can effectively handle various challenges in real scenarios.

D. Evaluation on RGBT210 Dataset

Fig. 9 shows the comparison results of our designed tracker with 10 advanced trackers on RGBT210 dataset. The comparison trackers are MANet [45], SGT [29], DSST+RGBT [69], MEEM+RGBT [68], SOWP+RGBT [67], KCF+RGBT [70], SOWP [67], CCOT [65], MDNet [63] and SiameseFC [4]. It is not hard to see that our designed tracker obtains the best performance against other advanced trackers in both evaluation metrics. Specifically, our designed tracker attains 77.9%/56.7% in PR/SR and achieves 2.7%/5.0% improvement over the second-best tracker MANet. Furthermore, compared with other competitive trackers, i.e., SGT and SOWP+RGBT, our designed tracker achieves 10.4%/13.8% and 13.5%/15.9% performance improvement in PR/SR, respectively. These evaluation results prove the feasibility of our designed tracker.

E. Evaluation on LasHeR Dataset

Fig. 10 shows the evaluation results of our designed tracker and 12 advanced trackers, including CMR [62], DAFNet [71], MANet [45], MANet++ [72], FANet [50], DAPNet [25], DMCNet [73], CAT [74], MaCNet [59], mfDiMP [46], SGT++ [30]

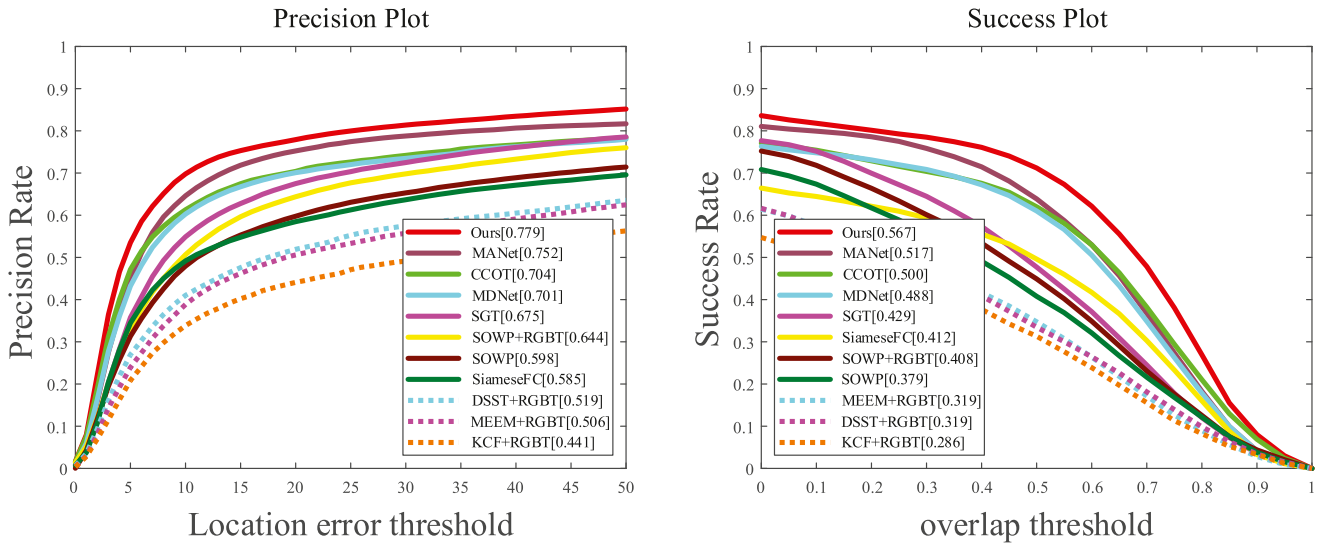

Fig. 9. Evaluation curves of different trackers on RGBT210 dataset.

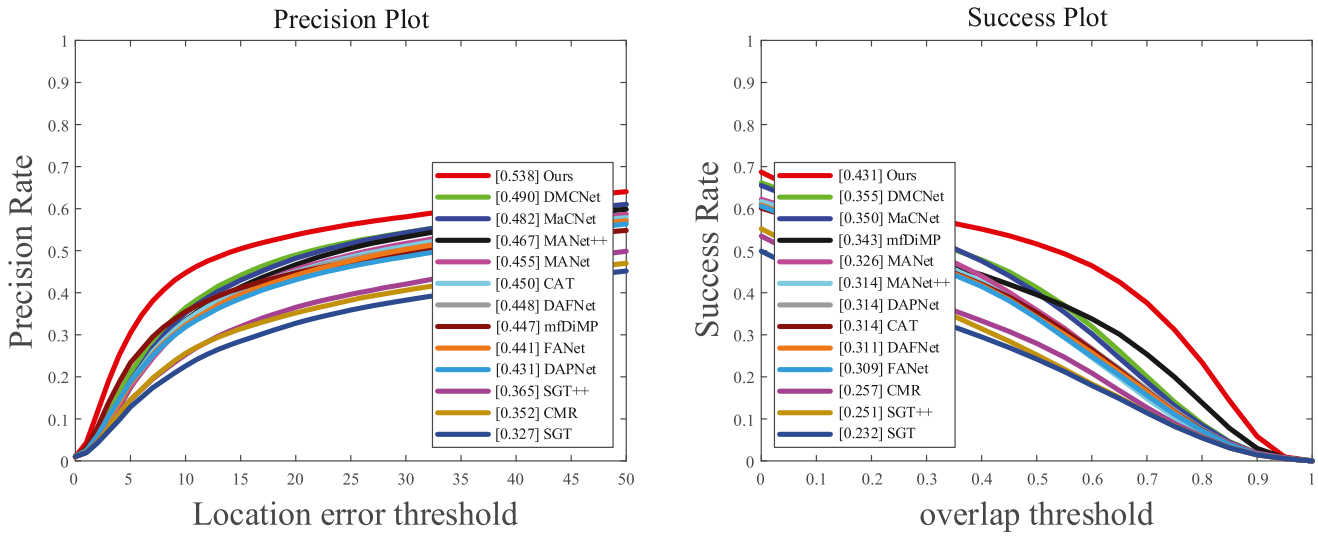

Fig. 10. Evaluation curves different trackers on LasHeR test dataset.

and SGT [29], on the LasHeR test dataset. It is not hard to see that our designed tracker achieves outstanding performance with 53.8%/43.1% in PR/SR. Specifically, our designed tracker outperforms DMCNet ranking second in PR/SR over 4.8%/7.6%. Compared with recent trackers mfDiMP and FANet, our designed tracker has 9.1%/8.8% and 9.7%/12.2% performance improvement, respectively. These results also fully prove the outstanding tracking ability of the designed tracker.

In addition, we evaluate the designed tracker against the retrained trackers MANet and mfDiMP on the LasHeR training set to further prove the effectiveness of the designed multi-layer attention aggregation Siamese network. Table II presents the evaluation results and we can see that our designed tracker still has more competitive performance. Although our designed tracker achieves 53.8%/43.1% in PR/SR and is slightly lower than retrained MANet, it runs 19 times faster than MANet. Compared with retrained mfDiMP, our designed tracker is 0.4% lower than it in PR, but our designed tracker has 6.3% higher in SR.

TABLE II
PR/SR VALUES (%) OF OUR DESIGNED TRACKER AND RETRAINED RGBT TRACKERS ON LASHER TEST DATASET

		MANet [32]	mfDiMP [33]	Ours
LasHeR	PR	60.7	54.2	53.8
testing set	SR	46.1	36.8	43.1

F. Ablation Study

Component analysis: To validate the contribution of the major components of our designed tracker, five variants are implemented on RGBT234, including: 1) Baseline, that extends the SiamCAR [28] into a dual-modality version for RGBT

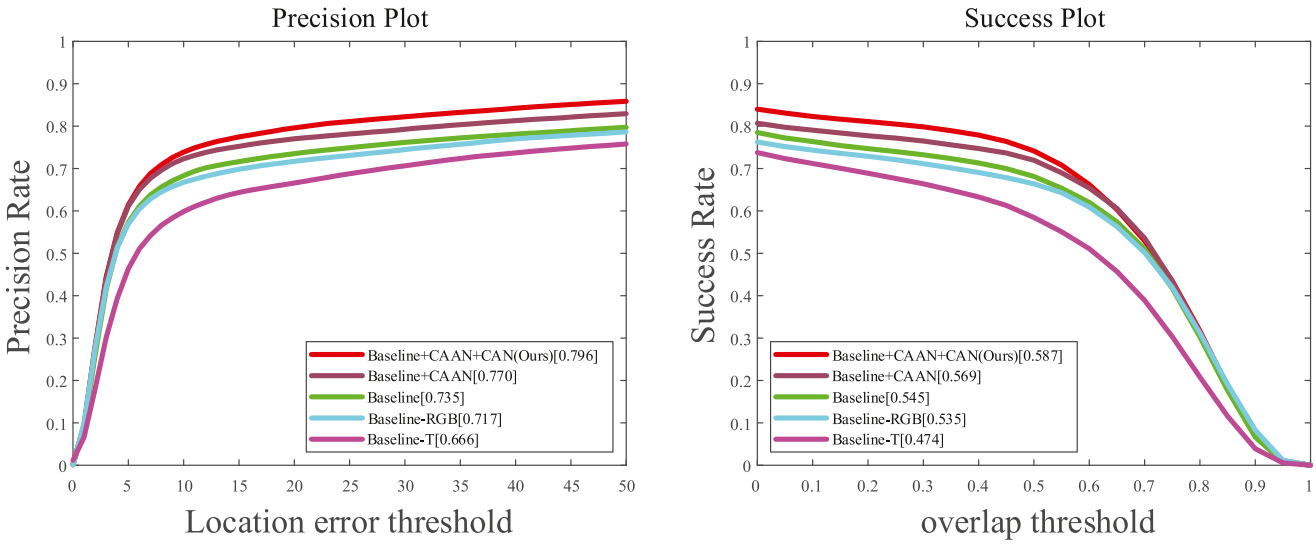

Fig. 11. Evaluation curves of our designed tracker with different variants on RGBT234 dataset.

tracking; 2) Baseline-RGB, that locates the target using only RGB modality; 3) Baseline-T, that locates the target using only thermal modality. 4) Baseline+CAAN, that integrates contribution-aware aggregation network in Baseline. 5) Baseline+CAAN+CAN, that incorporates the channel attention network in Baseline+CAAN. The comparison results of these versions are reported in Fig. 11.

From the result, it is easy to see that Baseline achieves 73.5%/54.5% in PR/SR and outperforms Baseline-RGB and Baseline-T. In particular, compared with Baseline-RGB and Baseline-T, Baseline has 1.8%/1.0% and 6.9%/7.1% improvement in PR/SR, respectively. This demonstrates the effectiveness of utilizing complementary RGB and thermal information for visual tracking. In addition, Baseline+CAAN attains 77.0%/56.9% in PR/SR and achieves 3.5%/2.4% promotion compared with Baseline, which validates the effectiveness of the contribution-aware aggregation network. Baseline+CAAN+CAN achieves 79.6%/58.7% in PR/SR and has 2.6%/1.8% promotion compared with Baseline+CAAN, which validates that the channel attention network is helpful for better target representation.

To more intuitively prove the effectiveness of the major components of our designed tracker, the partial quantized tracking results of Baseline, Baseline+CAAN, and Baseline+CAAN+CAN are presented in Fig. 12. In the sequence *child1* and *face1*, we can see that Baseline+CAAN can more accurately locate the target while Baseline fails to track the object in complex scenarios including light influence, heavy occlusion, and low illumination. These visual tracking results further prove the contribution-aware aggregation network is effective to learn the reliable modal weight. In the sequence *dog1* and *child3*, when the target faces the interference of background clutter and low resolution, Baseline+CAAN+CAN can successfully track the target while others lose it. This further proves that the channel attention network is able to learn powerful feature information for robust tracking. Overall, each component of our designed tracker is helpful to boost the tracking performance.

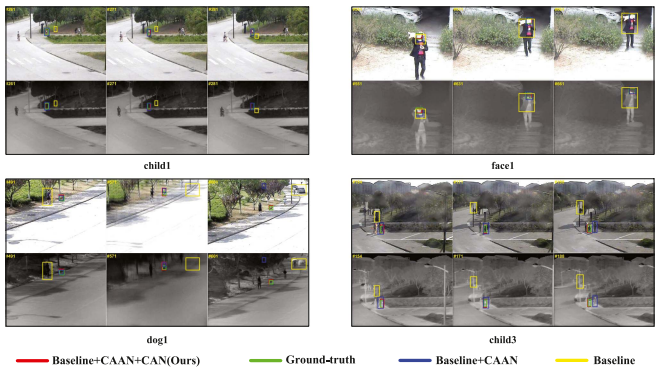

Fig. 12. Visual results of Baseline tracker, Baseline+CAAN tracker, and Baseline+CAAN+CAN tracker on four video sequences.

TABLE III
PR/SR VALUES (%) OF OUR DESIGNED TRACKER IN DIFFERENT FUSION LEVELS ON RGBT210 AND RGBT234 DATASETS

		Pix-level	Features-level	Decision-level
RGBT210	PR	59.4	63.2	**77.9**
	SR	43.9	44.8	**56.7**
RGBT234	PR	61.7	65.7	**79.5**
	SR	45.6	47.3	**58.4**

Analysis of different fusion levels: Some variants of our designed tracker in different fusion levels are implemented to prove the superiority of the decision-level fusion. These variants have three forms and are named Ours-Pix-Level, Ours-Feature-Level, and Ours-Decision-Level, respectively. Ours-Pix-Level directly concatenates two modal images to form 6-channel input information. Ours-Feature-Level concatenates the multi-modal feature maps extracted by the first convolutional layer. Ours-Decision-Level means the fusion level used by our tracker in the article. The comparison results of these versions on RGBT210 and RGBT234 benchmark datasets are reported in Table III. We

TABLE IV
PR/SR VALUES (%) SCORES OF OUR DESIGNED TRACKER WITH OTHER SIAMESE-BASED TRACKERS ON GTOT AND RGBT234 DATASETS

		SiamBAN [2]	SiamRPN++ [39]	DuSiamRT [47]	SiamFT [26]	SiamCDA [19]	Ours
GTOT	PR	71.7	72.5	76.6	75.8	87.7	**91.3**
	SR	59.3	61.7	62.8	62.3	73.2	**75.1**
RGBT234	PR	68.1	69.7	56.7	68.8	76.0	**79.5**
	SR	49.1	51.7	38.4	48.6	56.9	**58.4**

TABLE V
PR/SR VALUES (%) OF OUR DESIGNED TRACKER WITH THE ATTENTION-BASED TRACKERS ON GTOT, RGBT210, AND RGBT234 DATASETS. THE ✘ MEANS THAT THE TRACKER HAS NO RESULTS ON THE CORRESPONDING DATASET

	Tracker	DAFNet [71]	SiamFT [26]	DSiamMFT [27]	FANet [50]	M^{5}L [75]	HMFT [76]	Ours
GTOT	PR	89.1	82.6	✘	89.1	89.6	91.2	91.3
	SR	71.2	70.0	✘	72.8	71.0	74.9	75.1
RGBT210	PR	✘	✘	64.2	✘	✘	78.6	77.9
	SR	✘	✘	43.2	✘	✘	53.5	56.7
RGBT234	PR	79.6	68.8	✘	78.7	79.5	78.8	79.5
	SR	54.4	48.6	✘	55.3	54.2	56.8	58.4

The best, second, and third results are in red, green, and blue colors, respectively.

can see that Ours-Decision-Level is obviously superior to Ours-Pix-Level and Ours-Feature-Level. Specifically, Ours-Decision-Level obtains 14.7%/11.9% and 13.8%/11.1% performance improvement in PR/SR on RGBT210 and RGBT234 datasets over the feature-level fusion, which is the common fusion scheme used in RGBT tracking. These results prove the superiority of decision-level fusion in RGBT tracking.

Analysis of the Siamese-based RGBT trackers: We also evaluate the tracking performance of our designed tracker and several latest Siamese-based RGBT trackers on GTOT and RGBT234 datasets. The comparison Siamese-based trackers include SiamBAN [2], SiamRPN++ [39], DuSiamRT [47], SiamFT [26] and SiamCDA [19]. Table IV presents the evaluation results. It is not hard to see that our designed tracker performs best against all the compared Siamese-based trackers. Specifically, the performance of our designed tracker achieves 3.6%/1.9% and 3.5%/1.5% performance improvement in PR/SR against the most competitive Siamese-based RGBT tracker SiamCDA on GTOT and RGBT234, respectively. This proves the effectiveness of the designed multi-layer attention aggregation Siamese network. Furthermore, the significant performance advantages over other Siamese-based trackers, i.e., SiamFT and DuSiamRT indicate that our designed tracker can better exploit and utilize the complementary RGB and thermal information to enhance tracking ability.

Compare the RGBT trackers based on attention network: Table V presents the comparison results of our designed tracker with attention-based trackers, including DAFNet [71], SiamFT [26], DSiamMFT [27], FANet [50], M^{5}L [75], and HMFT [76] on GTOT, RGBT210, and RGBT234 datasets. It is not hard to see that our designed tracker performs best against these attention-based trackers in SR on all datasets. In particular, our designed tracker achieves 91.3%/75.1% and 79.5%/58.4% in PR/SR on GTOT and RGBT234, respectively. Compared with the most competitive attention-based tracker HMFT, our designed tracker obtains 0.1%/0.2% and 0.7%/1.6% performance improvement, respectively. Although our designed tracker is 0.7% lower in PR than HMFT on RGBT210, our designed tracker is 3.2% higher in SR than it. This demonstrates that the designed hierarchical interaction channel attention network can obtain more discriminative information for robust tracking. Furthermore, the significant performance advantages over other attention-based trackers, i.e., FANet and DSiamMFT verify that the proposed tracker can give full play to the complementary RGB and thermal information to accurately locate the target.

Compare the latest and most competitive RGBT trackers: To further prove the effectiveness of our designed tracker, some latest and most competitive trackers, including TFNet [18], ADRNet [77], APFNet [78], DMCNet [73], CMPP [79], and CAT [74] are added on GTOT, RGBT210, and RGBT234 for comparison. Table VI presents the evaluation results and our designed tracker exhibits competitive performance. To be specific, the SR scores of our designed tracker are 75.1% and 56.7% on GTOT and RGBT210, respectively, which outperforms all compared trackers in SR. Compared with TFNet, our designed tracker obtains 2.7%/2.2% and 0.2%/3.8% improvement in PR/SR on GTOT and RGBT210. Although TFNet outperforms our designed tracker by 1.1% in PR on RGBT234, it is 2.4% lower than our designed tracker in SR. Compared with CAT, our designed tracker has 2.4%/3.4% improvement in PR/SR on GTOT. On RGBT210 and RGBT234, CAT performs better over our designed tracker in PR, but it is much weaker in SR. Also, compared with CMPP, our designed tracker performs slightly worse in PR on GTOT and RGBT234, but it is much better in SR. Although DMCNet outperforms our designed tracker in PR/SR on RGBT234, our designed tracker performs much better than it on GTOT in both evaluation criteria. In addition, our designed tracker has 0.9%/1.2% improvement in PR/SR compared with

TABLE VI
PR/SR VALUES (%) OF OUR DESIGNED TRACKER WITH THE LATEST RGBT TRACKERS ON GTOT, RGBT210, AND RGBT234 DATASETS. THE ✘ MEANS THAT THE TRACKER HAS NO RESULTS ON THE CORRESPONDING DATASET

	Tracker	CMPP [79]	CAT [74]	ADRNet [77]	DMCNet [73]	APFNet [78]	TFNet [18]	Ours
	Pub.Info	CVPR2020	ECCV2020	IJCV2021	TNNLS2022	AAAI2022	TCSVT2022	
GTOT	PR	92.6	88.9	90.4	90.9	90.5	88.6	91.3
	SR	73.8	71.7	73.9	73.3	73.9	72.9	75.1
RGBT210	PR	✘	79.2	✘	79.7	✘	77.7	77.9
	SR	✘	53.3	✘	55.5	✘	52.9	56.7
RGBT234	PR	82.3	80.4	80.9	83.9	82.7	80.6	79.5
	SR	57.3	56.1	57.1	59.3	57.9	56.0	58.4

The best, second, and third results are in red, green, and blue colors, respectively.

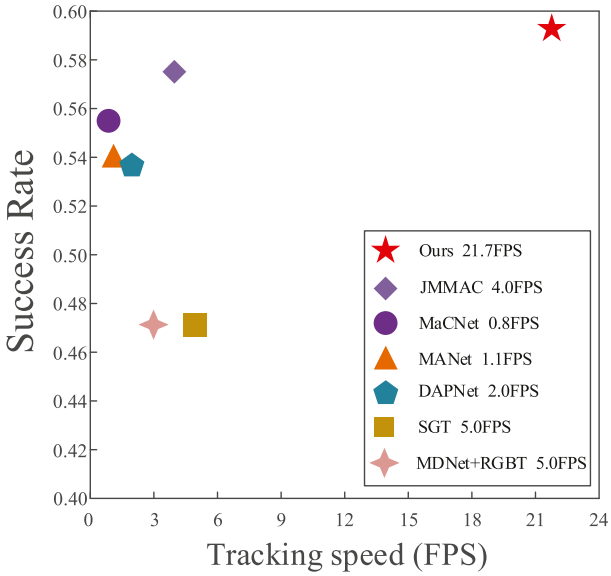

Fig. 13. Speed comparison of our designed tracker against other advanced trackers on RGBT234.

ADRNet on GTOT. Compared with APFNet, which is the latest and most competitive RGBT tracker, our designed tracker has 0.8%/1.2% improvement on GTOT. Although APFNet outperforms our designed tracker in PR on RGBT234, our designed tracker performs much better in SR.

G. Efficiency Analysis

Our designed tracker is implemented in Python using PyTorch on a server with Intel Xeon(R) E5-2620 CPU, 48 G RAM, Nvidia GTX 1080Ti. Fig. 13 shows the runtime of our designed tracker against other advanced RGBT trackers, including JMMAC [58], DAPNet [25], MaCNet [59], SGT [29], MANet [45] and MDNet+RGBT [63] on RGBT234 dataset. It is not hard to see that our designed tracker achieves faster tracking speed than that of the other trackers. Specifically, our designed tracker reaches 21.7FPS and is 16.7FPS faster than the second-best tracker SGT. In addition, compared with JMMAC and MaCNet, our designed tracker achieves 17.7FPS and 20.9FPS speed improvement, respectively. Overall, our designed tracker achieves superior performance with considerable speed.

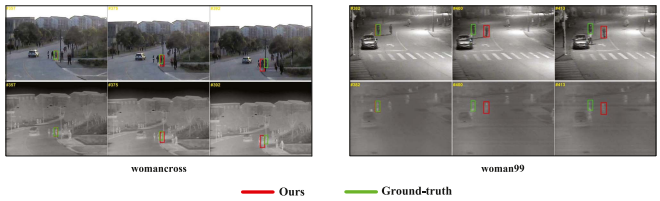

Fig. 14. Failure cases on *womancross* and *woman99* (RGBT234 dataset).

H. Failure Cases

Fig. 14 presents several cases where our designed tracker fails. In the sequence *womancross*, we can see that our tracker cannot accurately locate the target in the whole tracking process. The reason for the failure may be that suddenly moving camera leads to the tracking target falling outside the limited search range. Furthermore, intense camera shake easily leads to blurred images, making it difficult for our tracker to distinguish the target and background. In the sequence *woman99*, when the target suffers from heavy occlusion and thermal crossover, our tracker fails to determine the location of the target. The main reason is that the target is simultaneously disturbed by occlusion and similar targets, which easily leads to tracker drift. In the future, we will explore more advanced fusion structures in our framework to achieve more robust tracking.

V. CONCLUSION

In this article, a multi-layer attention aggregation Siamese network has been proposed for robust RGBT tracking. In particular, a channel attention network is incorporated into the Siamese network to recalibrate the feature channels of the multi-layer features, which can learn more robust feature representation. In addition, a contribution-aware aggregation network is designed to adaptively aggregate the responses of RGB and thermal branches, and thus can better utilize the complementary information between them. Lastly, a classification and regression network is introduced to determine the target state. Extensive experiments on four challenging RGBT benchmark datasets demonstrate outstanding performance against state-of-the-art trackers. In future work, we will explore more advanced fusion mechanisms to fully exploit the complementary strengths of different

modalities and investigate more suitable lightweight networks to further boost effectiveness and efficiency.

REFERENCES

[1] K. Yang et al., "SiamCorners: Siamese corner networks for visual tracking," *IEEE Trans. Multimedia*, vol. 24, pp. 1956–1967, 2022.

[2] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6668–6677.

[3] W. Ruan et al., "Multi-correlation filters with triangle-structure constraints for object tracking," *IEEE Trans. Multimedia*, vol. 21, pp. 1122–1134, 2019.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[5] X. Dong et al., "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[6] J. Shen et al., "Distilled siamese networks for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, Dec. 2022.

[7] X. Dong, J. Shen, L. Shao, and F. Porikli, "CLNet: A compact latent network for fast adjusting siamese trackers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 378–395.

[8] X. Dong et al., "Hyperparameter optimization for tracking with continuous deep q-learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 518–527.

[9] X. Dong et al., "Dynamical hyperparameter optimization via deep reinforcement learning in tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1515–1529, May 2021.

[10] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 459–474.

[11] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention Siamese network," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3068–3080, Jul. 2020.

[12] Z. Liang and J. Shen, "Local semantic Siamese networks for fast tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 3351–3364, 2020.

[13] Y. Zhang, B. Ma, J. Wu, L. Huang, and J. Shen, "Capturing relevant context for visual tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 4232–4244, 2021.

[14] X. Lu et al., "Deep object tracking with shrinkage loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2386–2401, May 2022.

[15] W. Han, X. Dong, F. S. Khan, L. Shao, and J. Shen, "Learning to fuse asymmetric feature maps in siamese trackers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16570–16580.

[16] S. Tian, X. Liu, M. Liu, S. Li, and B. Yin, "Siamese tracking network with informative enhanced loss," *IEEE Trans. Multimedia*, vol. 23, pp. 120–132, 2021.

[17] B. Jiang, Y. Zhang, B. Luo, X. Cao, and J. Tang, "STGL: Spatial-temporal graph representation and learning for visual tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2162–2171, 2021.

[18] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, "RGBT tracking by trident fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 579–592, Feb. 2022.

[19] T. Zhang, X. Liu, Q. Zhang, and J. Han, "SiamCDA: Complementarity- and distractor-aware RGB-T tracking based on siamese network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1403–1417, Mar. 2022.

[20] Q. Zhang et al., "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[21] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, May 2021.

[22] C. Li et al., "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.

[23] M. Feng, K. Song, Y. Wang, J. Liu, and Y. Yan, "Learning discriminative update adaptive spatial-temporal regularized correlation filter for RGB-T tracking," *J. Vis. Commun. Image Representation*, vol. 72, 2016, Art. no. 102881.

[24] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for RGB-T object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[25] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for RGBT tracking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 465–472.

[26] X. Zhang et al., "SiamFT: An RGB-Infrared fusion tracking method via fully convolutional siamese networks," *IEEE Access*, vol. 7, pp. 122122–122133, 2019.

[27] X. Zhang, P. Ye, S. Peng, J. Liu, and G. Xiao, "DSiamMFT: An RGB-T fusion tracking method via dynamic siamese networks using multi-layer feature fusion," *Signal Process.: Image Commun.*, vol. 84, 2020, Art. no. 115756.

[28] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6269–6277.

[29] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1856–1864.

[30] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106977.

[31] C. Li et al., "LasHeR: A large-scale high-diversity benchmark for RGBT tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 392–404, 2021.

[32] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4660–4669.

[33] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6182–6191.

[34] Q. Guo et al., "Learning dynamic siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1763–1771.

[35] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8971–8980.

[36] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.

[37] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448–1457.

[38] Z. Zhu et al., "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.

[39] B. Li et al., "SiamRPN : Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.

[40] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4591–4600.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Sci. China Inf. Sci.*, vol. 55, no. 3, pp. 590–599, 2012.

[43] S. Zhai, P. Shao, X. Liang, and X. Wang, "Fast RGB-T tracking via cross-modal correlation filters," *Neurocomputing*, vol. 334, pp. 172–181, 2019.

[44] C. Luo, B. Sun, K. Yang, T. Lu, and W.-C. Yeh, "Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme," *Infrared Phys. Technol.*, vol. 99, pp. 265–276, 2019.

[45] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2262–2270.

[46] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGBT-T tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2252–2261.

[47] C. Guo, D. Yang, C. Li, and P. Song, "Dual siamese network for RGBT tracking via fusing predicted position maps," *The Vis. Comput.*, vol. 38, no. 7, pp. 2555–2567, 2021.

[48] N. Cvejic et al., "The effect of pixel-level fusion on object tracking in multi-sensor surveillance video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[49] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation26 for robust visual tracking," in *Proc. IEEE 14th Int. Conf. Inf. Fusion*, 2011, pp. 1–8.

[50] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust RGBT tracking," *IEEE Trans. Intell. Veh.*, vol. 6, no. 1, pp. 121–130, Mar. 2021.

[51] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 664–679.
[52] L. Zheng et al., "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1741–1750.
[53] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005.
[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[55] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.
[56] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
[57] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[58] P. Zhang et al., "Jointly modeling motion and appearance cues for robust RGB-T tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3335–3347, 2021.
[59] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in RGB-T videos using modal-aware attention network and competitive learning," *Sensors*, vol. 20, no. 2, 2020, Art. no. 393.
[60] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 83–98.
[61] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1931–1941.
[62] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 808–823.
[63] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4293–4302.
[64] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6638–6646.
[65] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.
[66] J. Mei, D. Zhou, J. Cao, R. Nie, and Y. Guo, "HDINet: Hierarchical dual-sensor interaction network for RGBT tracking," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16915–16926, Aug. 2021.
[67] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "SOWP: Spatially ordered and weighted patch descriptor for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3011–3019.
[68] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.
[69] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.
[70] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with Kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[71] Y. Gao et al., "Deep adaptive fusion network for high performance RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 91–99.
[72] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5613–5625, 2021.
[73] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, "Duality-gated mutual condition network for RGBT tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar., 2022, doi: 10.1109/TNNLS.2022.3157594.
[74] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware RGBT tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 222–237.
[75] Z. Tu, C. Lin, W. Zhao, C. Li, and J. Tang, "M^{5}L: Multi-modal multi-margin metric learning for RGBT tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 85–98, 2022.
[76] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal UAV tracking: A. large-scale benchmark and new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8886–8895.
[77] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time RGB-T tracking," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2714–2729, 2021.
[78] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for RGBT tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2831–2838.
[79] C. Wang et al., "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7064–7073.

Mingzheng Feng received the B.S. degree from Yanshan University, Qinhuangdao, China, in 2018, and the M.S. degree from Northeast University, Shenyang, China, in 2021. He is currently working toward the Ph.D. degree with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision, deep learning, and multimodal fusion.

Jianbo Su (Senior Member, IEEE) received the B.S. degree in automatic control from Shanghai Jiao Tong University, Shanghai, China, in 1989, the M.S. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Science, Beijing, China, in 1992, and the Ph.D. degree in control science and engineering from Southeast University, Nanjing, China, in 1995. In 1997, he joined the Faculty of the Department of Automation, Shanghai Jiao Tong University, where he has been a Full Professor since 2000. In his research areas, he has authored or coauthored three books, more than 300 technical papers. and is the holder of 25 patents. His research interests include robotics, pattern recognition, and human-machine interaction.