Knowledge-Based Systems 249 (2022) 108945

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Learning reliable modal weight with transformer for robust RGBT tracking

Mingzheng Feng, Jianbo Su *

Department of Automation, Shanghai Jiao Tong University and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 20 January 2022
Received in revised form 26 April 2022
Accepted 27 April 2022
Available online 11 May 2022

Keywords:
RGBT tracking
Transformer
Semantic features

ABSTRACT

Many Siamese-based RGBT trackers have been prevalently designed in recent years for fast-tracking. However, the correlation operation in them is a local linear matching process, which may easily lose semantic information required inevitably by high-precision trackers. In this paper, we propose a strong cross-modal model based on transformer for robust RGBT tracking. Specifically, a simple dual-flow convolutional network is designed to extract and fuse dual-modal features, with comparably lower complexity. Besides, to enhance the feature representation and deepen semantic features, a modal weight allocation strategy and a backbone feature extracted network based on modified Resnet-50 are designed, respectively. Also, an attention-based transformer feature fusion network is adopted to improve long-distance feature association to decrease the loss of semantic information. Finally, a classification regression subnetwork is investigated to accurately predict the state of the target. Sufficient experiments have been implemented on the RGBT234, RGBT210, GTOT and LasHeR datasets, demonstrating more outstanding tracking performance against the state-of-the-art RGBT trackers.

1. Introduction

Visual tracking [1–3] has attracted ascendant attention for its versatile applications in automatic driving, intelligent surveillance, and robot navigation. Visual tracking is committed to estimating the location and scale of a specific object in subsequent frames, given its state in the first frame. Although visual tracking has achieved great success with the help of the robust object representation brought by deep neural networks, the robustness of these trackers [4,5] still needs further improvement in challenging scenes such as fast motion, low resolution, and illumination change.

There have been many studies on the robustness of trackers in complex challenging scenarios by applying multi-modal information. Recently, the RGBT tracking [6,7], which can integrate the advantages of visible and thermal infrared information, has widely been investigated. Specifically, compared with visible image, thermal infrared image has strong penetration ability and is not sensitive to illumination change [8,9]. It can accurately capture the object under extreme weather conditions, such as night, fog, or haze, as can be seen in Fig. 1(a). Nevertheless, when encountering the challenge of the thermal cross, the thermal infrared image tends to confuse different objects. On the other hand, the visible image can distinguish more detailed information such as color and texture, and has a stronger resolution in forest-background separation, as can be seen in Fig. 1(b). Thus, the RGBT tracking can preferably deal with complex challenging scenarios.

Many RGBT trackers have been proposed so far. Early RGBT trackers use sparse learning-based method to achieve dual-modal information fusion. Liu et al. [10] construct the likelihood function of the particle filter tracking algorithm and realized the final fusion of two modal information by minimizing the joint sparse representation coefficient. Li et al. [11] integrate collaborative sparse representation and modal weights in the Bayesian framework to fuse two modal information. However, these trackers are prone to tracking failure when the sparse representation or classification score is insufficient to reflect modal reliability. In recent years, some excellent RGBT trackers have emerged with the extensive explorations and applications of the correlation filter. Zhai et al. [12] jointly learn different modal correlation filters by using low-rank constraint and then achieve the consistent positioning of the object. Feng et al. [13] learn an adaptive spatial–temporal regularized coefficient to build model and design a weighted ensemble strategy to integrate the information between RGB and thermal infrared. However, these methods only use the object features extracted by hand-crafted and cannot represent the object information well. Inspired by widely application of convolutional neural network (CNN) in RGB tracking, Li et al. [7] design a dual-stream ConvNet and a FusionNet, which can get rich semantic information in deep layers and complete

* Corresponding author.
E-mail address: jbsu@sjtu.edu.cn (J. Su).

https://doi.org/10.1016/j.knosys.2022.108945
0950-7051/

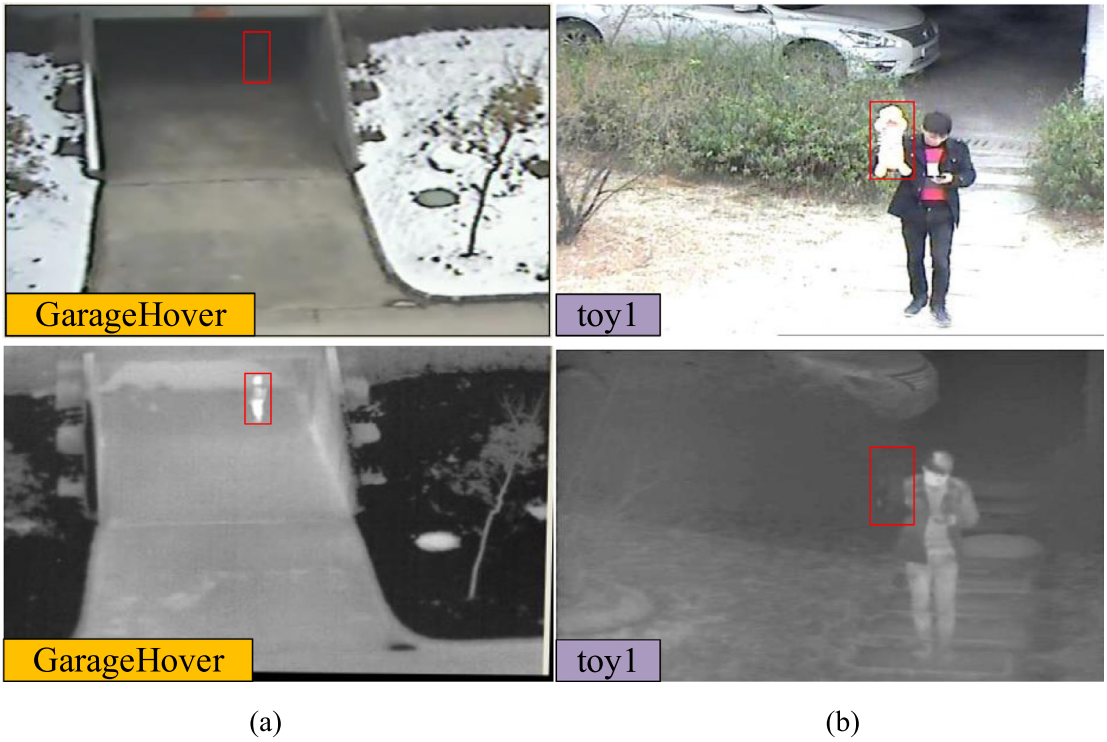

Fig. 1. Illustration of the tracking advantage between multi-modal images. (a) The advantage of thermal infrared modality over RGB modality. (b) The advantage of RGB modality over thermal infrared modality. As can be seen from the red box, the complementary advantage between the two modal images is obvious.

adaptive fusion of different images information, but its tracking speed is far away from the real-time applications.

Inspired by the speed advantage of the Siamese network in RGB tracking, many researchers have explored the incorporation of Siamese networks to accelerate the RGBT trackers. Zhang et al. [14] first apply a fully convolution Siamese network in RGBT tracking and the tracker can run at about 30 frames per second. However, the correlation operation of Siamese-based trackers is a simple local linear matching process between template and search area. This process only uses the local information of the search area without considering the important proportion of tracking target in global information, easily falling into the local optimum. In addition, it tends to fragment the complete semantic information of the tracking target, which may result in difficulties of determining the target boundaries.

In this paper, we propose a novel end-to-end trainable cross-modal tracker based on transformer for robust RGBT tracking. Firstly, we design a simple dual-flow convolutional network to extract features from RGB and thermal infrared images respectively and then concatenate them. Secondly, a modal weight allocation strategy is designed to update the fused features information, which can enhance the resolution of fused features and effectively reduce the difference between modal features. Then, a backbone network based on modified Resnet-50 is used to extract deeper semantic features. After that, these features are reshaped into feature vectors and fed into a transformer feature fusion network to combine the template and search region features. Finally, the learned feature vectors are fed into the classification and regression network, and then complete the state estimation of the target. The main contributions can be summarized as follows:

- An RGBT tracking framework based on the transformer is designed, which can enhance long-distance feature association and decrease the loss of semantic information. To our knowledge, this is the first time to incorporate the transformer in RGBT tracking.
- A shallow convolutional network is designed to extract and fuse multi-modal information, which significantly simplifies the calculation process. Moreover, an optimal modal weight allocation strategy is proposed to obtain reliable weight for effectively optimizing fused features.
- A classification and regression subnetwork by adding a central branch is adopted to reduce the interference of background, further improving the accuracy of target prediction.
- Sufficient experimental results on four large benchmark datasets, RGBT234 [15], RGBT210 [16], GTOT [11] and LasHeR [17] indicate that the proposed tracker obtains more outstanding performance compared to the state-of-the-art RGBT trackers.

2. Related work

Visual tracking is regarded as one of the most fundamental computer vision tasks and has attracted more and more studies with more and more applications. This section will give a brief introduction on RGB tracking, RGBT tracking, and transformer mechanism.

2.1. RGB tracking

With the strong feature representation ability of the convolutional neural network, the RGB tracking based on deep learning has gradually become the mainstream method. Danelljan et al. [18] design a novel tracking architecture, capable of fully exploiting explicit components for target prediction. Bhat et al. [19] develop an end-to-end discriminative model prediction architecture for robust tracking. Also, there are several excellent TIR trackers based on deep learning. Liu et al. [20] design a multi-layer convolutional framework to extract rich information for object tracking. Liu et al. [8] regard tracking as a matching problem and train a matching network offline for online tracking. However, these trackers are limited by the weak discriminative

capacity of the learned features. To address this problem, Liu et al. [21] introduce a dual-level feature model to effectively distinguish distractors for robust object tracking. Currently, deep learning-based trackers mostly use the structure of Siamese networks because they are leading the way in the performance of popular tracking benchmarks. As one of the pioneering works, Bertinetto et al. [22] construct a fully convolution Siamese network to train the tracker and express the visual task as a similar learning problem. Based on [22], Valmadre et al. [23] integrate correlation filtering with basic feature representation by an online training scheme. Wang et al. [24] add an attention mechanism to improve tracking performance. However, the similarity measurement cannot cope with the scale change of the target based on deep network learning alone. Some researchers try to utilize a regression network to directly predict the target location. Li et al. [25] introduce a regional regression network, which can not only directly track target position through network regression, but also evaluate the confidence of each candidate region. Zhu et al. [26] further improve the semantic perception ability of the network model through data enhancement strategy and overcome the problem of an unbalanced distribution of training data. However, the performance of these trackers cannot consequently improve when deeper networks are used as the backbone. Li et al. [27] move the position of the object randomly and accordingly incorporate the Siamese network into the tracker. Zhang et al. [28] design a residual network to solve the problem that deep networks destroy strict translational invariance. Chen et al. [3] introduce a simple yet effective anchor-free Siamese framework to avoid the intricate parameters of anchor setting. To improve the tracking performance under interference scenarios, Cheng et al. [29] design a refinement module that can optimize the classification and regression branch to obtain a robust learning ability. These improvements enable the resultant trackers based on Siamese network to achieve better accuracy by using deeper network architecture.

2.2. RGBT tracking

With the popularity of thermal infrared sensors and the proposal of RGBT210 [16] and RGBT234 [15] tracking benchmarks, RGBT tracking has attracted extensive attention. Early work employs sparse representation for RGBT tracking. Wu et al. [30] concatenate multi-source data information into a one-dimensional vector and sparsely represent them for robust tracking. Li et al. [16] build a regularized map by adopting a weighted sparse representation to obtain strong information for visual tracking. However, these efforts will tend to fail when reconstruction residual cannot reliably calculate modal reliability. Recently, the correlation filter trackers have been widely designed because of their significant computational efficiency. Yun et al. [31] build a fusion model based on correlation filter and can discriminatively fuse different modal features. Luo et al. [32] exploit the multi-source information based on a tracking-by-detection architecture and designed an adaptive weighting scheme to fuse multi-modal information. Xu et al. [33] design novel multi-fusion levels to effectively integrate the information of RGB and thermal infrared images. Deep learning techniques have been widely used in many vision tasks. For RGBT object detection, Zhang et al. [34] design a novel end-to-end CNN network to solve the challenge of RGBT saliency detection. However, the reliability of different modalities is ignored. Zhang et al. [35] revisit different fusion strategies and design a novel RGBT fusion network to learn the importance of each modality. There are also many RGBT trackers based on deep learning. Lu et al. [36] design a multi-adapter network to obtain powerful RGBT representation. Wang et al. [37] design a modality-aware filter generation module, which can adaptively adjust the convolution kernels of different input images to enhance the communication between them. In addition, the RGBT tracking method based on the Siamese network has been widely used for its excellent performance. Zhang et al. [38] propose a multi-layer fusion tracking method based on dynamic Siamese network, which made the tracking process more robust. A complementarity-aware module is reported in [39] to enhance the discriminability of the fused features with increased accuracy. Although these RGBT trackers promote the development of the RGBT tracking, they tend to overlook the nature of feature interaction in the learning process, which may limit the further promotion of tracking performance.

2.3. Transformer mechanism

Transformer architecture is first proposed in [40] for machine translation tasks and has attracted much attention for its simple framework with excellent performance. Transformer architecture consists of the attention mechanism, which reduces the distance between any two positions in the input sequence to a constant and calculates the importance of each position with the rest of the sequence. Compared with RNN, the transformer is more competitive in long sequence tasks by its parallel computation and unique position encoding. In addition, it abandons traditional recursion and convolution, thus having less training time. The BERT algorithm [41] based on transformer has achieved excellent performance in NLP multi-tasks.

Recently, many researchers have tried to introduce transformer architecture into the field of computer vision and achieved excellent results in many fields. Zheng et al. [42] propose a variant of the transformer to improve the computational efficiency in object detection of high-resolution. Also, some efforts have been made to introduce transformer into visual tracking. Choi et al. [43] design a deep network based on attentional mechanism to adaptively choose the subset of the associated correlation filters for robust tracking. Yu et al. [44] propose a deformable attention network to enhance discrimination. However, these trackers are dependent on feature correlation operations on the fusion of template and searched areas. Hence, an attention-based fusion network is presented to combine the features of template and searched area in [45], which abandons the process of correlation operation. This work initiates our efforts to incorporate the transformer into RGBT tracker.

3. The transformer-based RGBT tracker

This section will first give an overview of the proposed RGBT tracker and then describe the detail of each component of the overall network architecture. The training loss of the proposed tracker is also described accordingly.

3.1. Overview

The overall flow chart of the proposed tracker is shown in Fig. 2 and we can find that the whole tracking process can be divided into four parts: shallow information fusion and weight optimization, deep semantic feature extraction network, transformer feature fusion network, classification and regression subnetwork.

Specifically, the clipped RGB and thermal infrared images are sent into the shallow dual-flow convolutional neural network for feature extraction. The extracted RGB features are then concatenated to the extracted thermal infrared features. Then a weight optimization strategy is invented to optimize the representation of the fused features. To obtain the rich semantic information of images, a backbone network by a modified Resnet-50 is adopted to obtain further semantic features. Then the extracted features

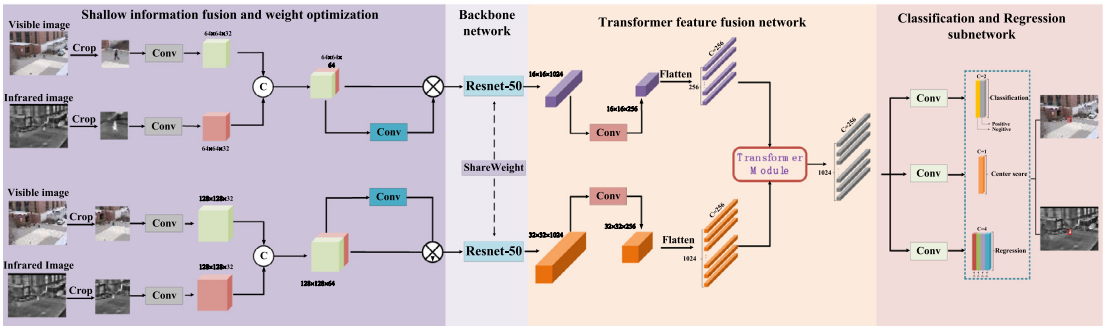

Fig. 2. The pipeline of the proposed RGBT tracker. The overall framework consists of four main parts: shallow information fusion and weight optimization, backbone framework for deeper semantic feature extraction, transformer feature fusion network for the fusion of template and search branches, classification and regression subnetwork for the prediction of target state.

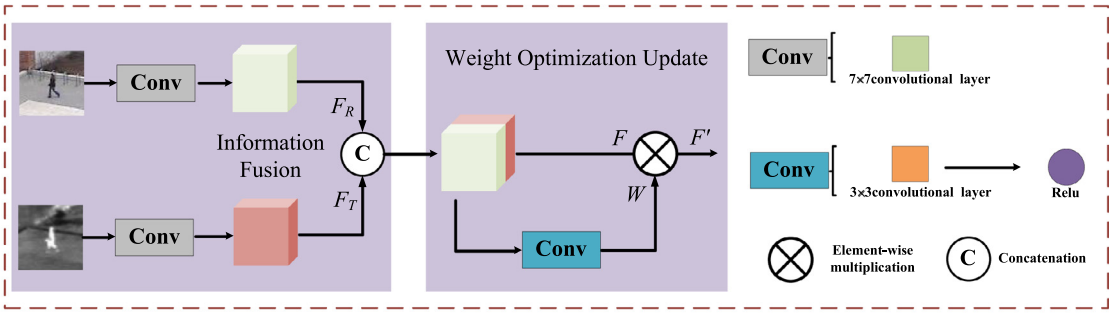

Fig. 3. Illustration of the shallow information fusion and weight optimization network. First, the features of RGB and thermal infrared images extracted by a simple dual-flow convolutional network are concatenated. Then the designed modal weight allocation strategy is to optimize the fused features, which can effectively enhance the resolution and reduce the difference of fused features.

are sent to the fusion network through flatten operation. Finally, the output feature vectors pass through the classification and regression subnetwork to complete the prediction of the target state.

3.2. Shallow information fusion and weight optimization

In recent years, RGBT tracking has attracted extensive attention for its unique advantages, and how to effectively fuse the complementary information of them to complete the desired tracking task is still the most essential problem. Many trackers cannot take full advantage of the complementary information from the difference among variant modal features. In addition, the fusion stage follows mostly that of the deep feature extraction, which may greatly reduce the efficiency of the tracing process. Hence, we design a shallow information fusion and weight optimization network, as shown in Fig. 3.

The input RGB and thermal infrared images are first extracted by a simple dual-flow convolutional network to obtain features F_R and F_T. The fused feature F is obtained by:

$$F = cat(F_R, F_T) \tag{1}$$

where $cat(*)$ defines the concatenation operation. To get the weight w, we make the fused feature F through the 3*3 convolution layer and sigmoid function:

$$w = \sigma(conv(F, \beta)) \tag{2}$$

where $conv(*, \beta)$ defines the convolution layer with parameters, σ defines the sigmoid function layer. The reliability of the modal information can be reflected by the generated weight. After getting the weight w and F, we can obtain:

$$\hat{F} = F * w \tag{3}$$

where $*$ represents the multiplication calculated by the element, and the optimization of feature representation is completed by weight optimization strategy.

3.3. Deep semantic feature extraction backbone network

To extract deeper features and rich semantic information, we use the modified Resnet-50 network as our backbone extraction network. As shown in Fig. 4, the first and the final stage of normal ResNet-50 are removed so that the output of the fourth stage is taken as the final output. In addition, to obtain a better feature resolution, we set the down-sampled convolution step of the fourth stage as 1.

3.4. Transformer feature fusion network

The fusion network of the baseline tracker Transt-tracking [45] is adopted here to arrive at a specific feature fusion process shown in Fig. 5.

From Fig. 5, we can see that the template feature F_t and search feature F_s which are extracted by the backbone network first pass through the 1*1 convolutional network and obtain $F_{t0} \in R^{d\times H_z\times W_z}$ and $F_{s0} \in R^{d\times H_x\times W_x}$. We flat the feature F_{t0} and F_{s0} in the spatial dimension, and get $F_{t1} \in R^{d\times H_z\times W_z}$ and $F_{s1} \in R^{d\times H_x\times W_x}$. Then the feature vectors $F_{t1} \in R^{d\times H_z\times W_z}$ and $F_{s1} \in R^{d\times H_x\times W_x}$, are fed into the TM module, as shown in Fig. 5. First, they through a transformer self-attentional module (TS). Then the transformer cross-attention module is designed to fuse different branch information. To be more specific, two transformer cross-attention modules are used to get the feature vectors of their own and the other branch and fuse them. To make the fusion information more accurate, the fusion process is repeated four times. Finally, an extra transformer cross-module is added to fuse the feature

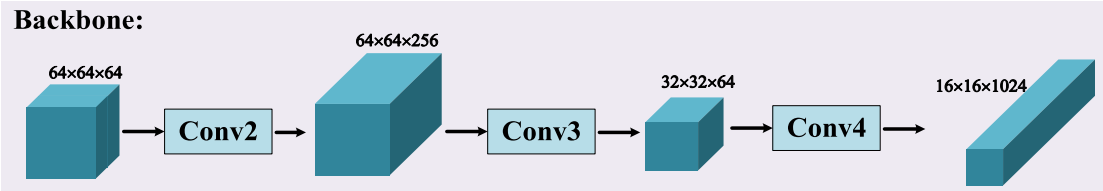

Fig. 4. Illustration of the backbone network constituted of the modified Resnet-50 network.

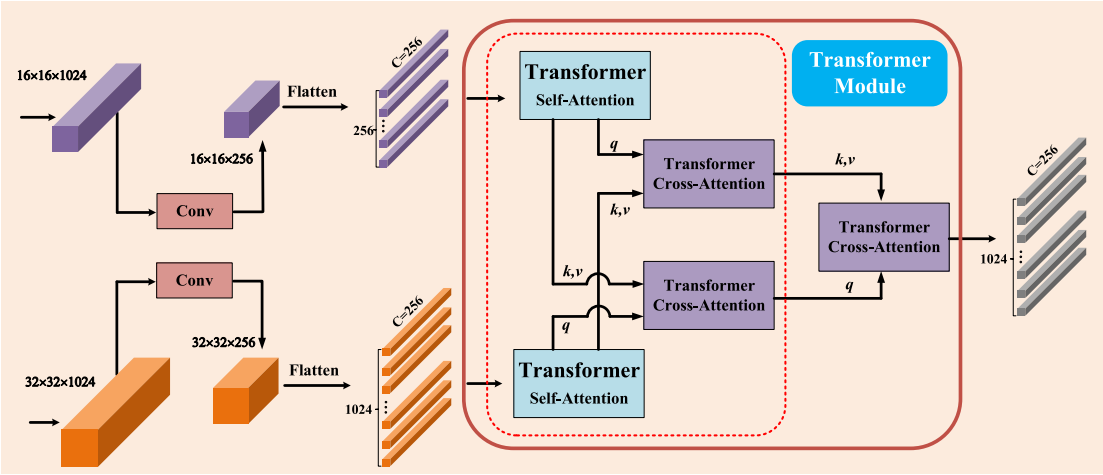

Fig. 5. Illustration of the features fusion network based on the transformer.

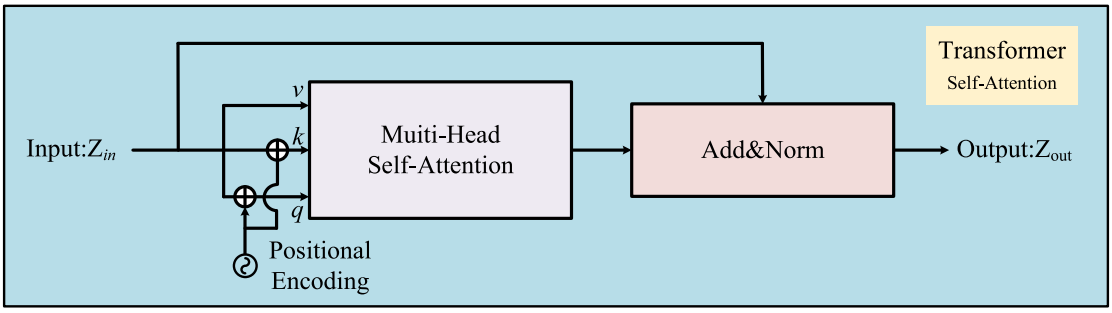

Fig. 6. Illustration of the transformer self-attention module. The module is mainly composed of the multi-head self-attention in a residual form.

vectors of the template and search branch. Next, we will give a brief introduction to TS and TC modules.

Fig. 6 shows the self-attention modules for transformer (TS). The TS module first introduces a positional encoding process to effectively distinguish the position information of feature sequences. Then the multi-head self-attention is used to integrate the feature vectors of different positions. Finally, residual form is used to get the output. The specific calculation process of the TS module is as follows:

$$Z_{EC} = Z + MultiHead(Z + P_x, Z + P_x, Z) \tag{4}$$

where $P_x \in R^{d\times N_x}$ represents the spatial positional encoding generated by using a sine function. $Z \in R^{d\times N_x}$ and $Z_{EC} \in R^{d\times N_x}$ is the input and the output of the TS module, respectively.

Fig. 7 shows the cross attention module for transformer (TC). The TC module first introduces a positional encoding process to effectively distinguish the position information of feature sequences. Then the multi-head cross-attention is used in residual form to integrate the feature vectors from different inputs. In addition, a feedforward network (FEN) is used in the form of residual to get the final output. The specific calculation process of the TC module is as follows:

$$\widetilde{Z}_{CF} = Z_q + MultiHead(Z_q + P_q, Z_{kv} + Z_{kv}, Z_{kv}) \tag{5}$$

$$Z_{CF} = \widetilde{Z}_{CF} + FEN(\widetilde{Z}_{CF}) \tag{6}$$

where $Z \in R^{d\times N_x}$ and $Z_{kv} \in R^{d\times N_{kv}}$ are the two inputs from different branches. $P_q \in R^{d\times N_q}$ and $P_{kv} \in R^{d\times N_{kv}}$ are the spatial positional encoding of Z_q and Z_{kv}, respectively. $\widetilde{Z}_{CF}$ and Z_{CF} are the output of the residual multi-head cross-attention and the final output, respectively.

3.5. Prediction with classification and regression subnetwork

As shown in Fig. 2, the subnetwork of classification and regression is composed of a classification branch, a regression branch, and a center-ness branch. The classification branch determines the location of the target by calculating the classification results of positive and negative samples; The regression branch abandons anchor boxes based on prior knowledge and directly predicts normalized coordinates to simplify the tracking framework. In addition, the locations far from the target center tend to produce

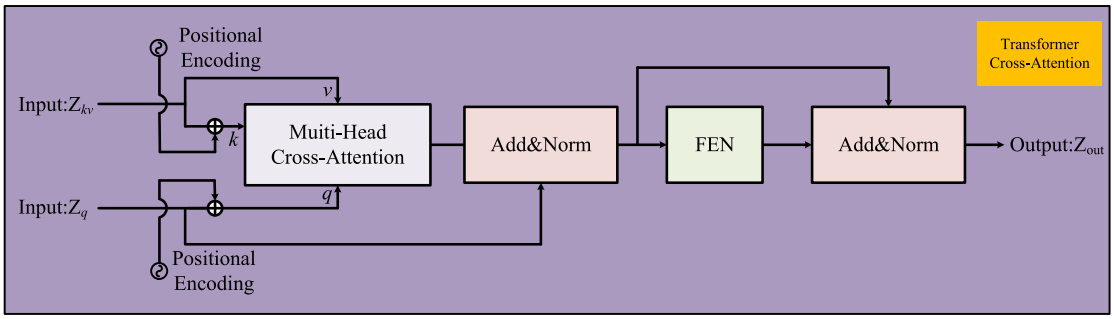

Fig. 7. Illustration of the transformer cross-attention module. The module is mainly composed of multi-head cross-attention and FFN in a residual form.

low-quality predictive bounding boxes. A central branch is therefore added to remove outliers, further improving the accuracy of target state prediction.

3.6. Training loss

The target on each search patch is marked as a ground real boundary box. We build the positive samples by selecting the predicted feature vector corresponding to the pixels in the ground-truth box. The negative samples are composed of the rest samples. The classification loss is decided by all samples and the regression loss is determined by positive samples.

The standard binary cross-entropy loss is adopted to measure the loss of classification:

$$L_{cls} = \sum_j [y_j \log(m_j) + (1 - y_j) \log(1 - m_j)] \tag{7}$$

where y_j defines the ground-truth label of the jth sample, $y_j = 1$ denotes foreground, and m_j represents the probability in the foreground.

Meanwhile, the linear combination of $l1$-norm loss $L_1(.,.)$ is employed to calculate the regression loss as follows:

$$L_{reg} = \sum_j \Pi\{y_j = 1\}[\lambda_G L_{GIOU}(b_j, \hat{b}) + \lambda_1 L_1(b_j, \hat{b})] \tag{8}$$

where $y_j = 1$ defines the positive sample, b_j denotes the jth output box, and $\hat{b}$ defines the normalized ground-truth box. The regularization parameters λ_G and λ_1 are set 2 and 5, respectively. $L_{GIOU}(.,.)$ denotes the generalized IoU loss.

The center-ness loss can be formulated by employing the BCEWithLogitsLoss function:

$$L_{cen} = -\sum_j [c_j \log \sigma(t_j) + (1 - c_j) \log(1 - \sigma(t_j))] \tag{9}$$

where t_j defines the jth predicted center-ness score of the corresponding location and σ defines the sigmoid function. c_j defines the percentage of the relative distance between the corresponding location and center location. If the corresponding location is not in the foreground, the value of c_j is set to 0.

The overall training loss can be calculated as follows:

$$L = \eta_1 L_{cls} + \eta_2 L_{reg} + \eta_3 L_{cen} \tag{10}$$

where η_1, η_2, η_3 represent the weight coefficients of classification loss, regression loss and center-ness loss, respectively.

4. Experiments

This section will first introduce the experiment setup and then compare the experimental results of the proposed tracker with those of the state-of-the-art ones on four public RGBT datasets, RGBT234 [15], RGBT210 [16], GTOT [11] and LasHeR [17]. In addition, the ablative study is given to further evaluate each component of the proposed tracker.

4.1. Experimental setup

4.1.1. Implementation details

The proposed RGBT tracker is trained on the LasHeR dataset [17] and the AdamW algorithm with the learning rate of 0.0001 is adopted to optimize the model. In addition, the values of weight decay and momentum are selected as 0.0001 and 0.9, respectively. The size of the cropped template patch and search region is 128 pixels and 256 pixels, respectively. The backbone network is composed of the modified ResNet-50 and the correspondent network parameters are initialized by the pre-trained model on the ImageNet dataset [46]. In the training process, the batch size is set to be 20, and 60 epochs with 2000 iterations per epoch are performed. In the process of calculating overall loss, the weight coefficients are set as $\eta_1 = 8.2$, $\eta_2 = 4.7$, $\eta_3 = 1.1$, respectively. For the tracking progress, the first frame of sequential images is used as the template patch. The search region in the current frame is regarded as the input of the search branch. In the classification-regression subnetwork, the box with the best score is regarded as the final output result. The proposed tracker is implemented based on Python 3.6, PyTorch 1.4.1, and all the experiments are run on a machine with CPU E5-2620 and four Nvidia GTX 1080Ti GPUs.

4.1.2. Experimental benchmarks

Four public RGBT datasets, including GTOT, RGBT210, RGBT234 and LasHeR are adopted to evaluate the performance of different trackers. A brief description of them is given as follows.

RGBT210 dataset is a large dataset used to evaluate the RGBT tracking methods. It contains 210 sets of RGB and thermal infrared video pairs with high precision alignment. The total number of video frames reaches 210 K, and all tracking targets are accurately marked with ground truth values. RGBT210 video sequences are divided into 12 different attributes such as low illumination and thermal crossover, to analyze attribute-based tracking.

RGBT234 is an extension of the RGBT210 dataset. It contains 234 highly aligned RGB and thermal infrared sequence pairs. The total frame number of RGBT234 is about 234 K, of which the longest sequence has about 8000 frames. As with RGBT210, the RGBT234 dataset also marks 12 attributes which are to evaluate the performance of tracking algorithms based on different attributes.

GTOT dataset is a standard benchmark dataset for RGBT tracking. It consists of 50 pairs of visible and thermal infrared video sequences with a total frame size of about 15 K. It can be divided into 7 subsets according to annotated attributes and all tracking targets are accurately marked with ground truths.

LasHeR dataset is a recently released benchmark for RGBT tracking. The dataset contains more than 734.8 K manually annotated frames and 1224 pairs of highly aligned visible and thermal infrared video sequences. LasHeR can be split into training and

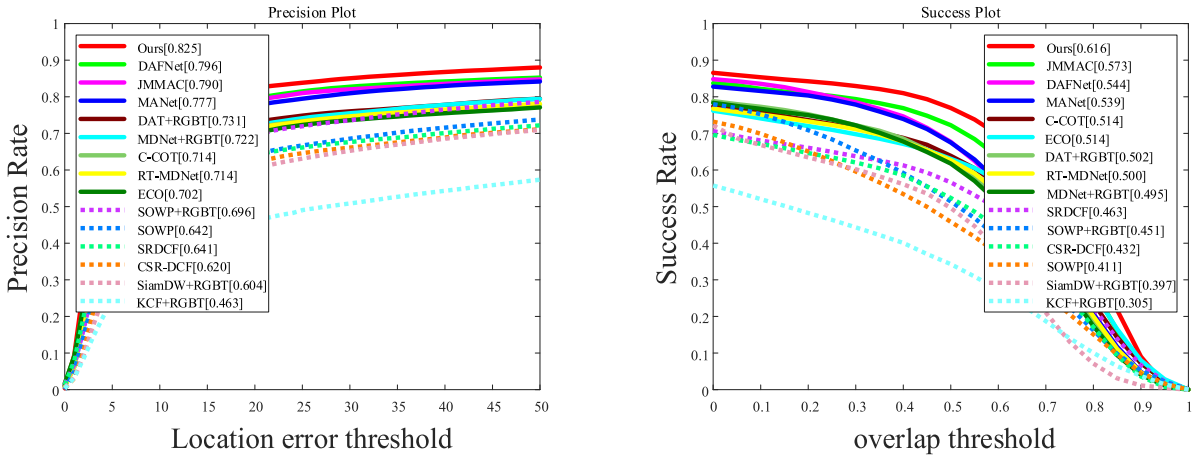

Fig. 8. Evaluation plots of precision and success on the RGBT234 dataset.

testing subsets according to the target class distribution. Such a large test dataset brings a great challenge to the tracking algorithms. Moreover, the LasHeR video sequences are divided into 19 different attributes such as frame lost and similar appearance, which make the tracking challenging.

4.1.3. Evaluation metrics

The well-defined two commonly used criteria, Precision Rate (PR) and Success Rate (SR), are employed to quantitatively evaluate the performance of each tracker.

PR is the percentage of frames whose distance between the output position and ground truth is within a given threshold. The given threshold is set differently in different datasets because of the target characteristics they contained. For the RGBT210 and RGBT234 datasets, the given threshold of distance is set to 20 pixels.

SR is the percentage of frames whose overlap ratio between the output box and the ground truth box is greater than a given threshold. Assuming that the target boundary box of the predicted output is R_P and the real target boundary box of manual annotation is R_G, the intersection ratio of the two boundary boxes is defined as:

$$S = \frac{|R_P \cap R_G|}{|R_P \cup R_G|} \tag{11}$$

where $\cap$ represents the intersection, $\cup$ represents the union, $|\cdot|$ represents the number of pixels contained. The area under the curve is employed to calculate the SR score in this paper.

4.2. Evaluation on RGBT234 dataset

Experimental results of the proposed tracker and other RGBT trackers on RGBT234 dataset are firstly reported. The compared trackers include DAFNet [47], RT-MDNet [48], DAT+RGBT [49], MDNet+RGBT [50], JMMAC [51], SOWP+RGBT [6], SRDCF [52], ECO [53], CSR-DCF [54], C–COT [55], KCF+RGBT [56], SOWP [6], MANet [57], SiamDW+RGBT [28]. Then we analyze the performance of these trackers on each attribute. Finally, the visualized result of different trackers is given to qualitatively verify the effectiveness of the proposed tracker.

4.2.1. The overall performance

As shown in Fig. 8, the proposed tracker reaches 82.5%/61.6% in PR/SR on the RGBT234 dataset and achieves the best performance among all trackers. More specifically, compared with the strong tracker JMMAC, which ranks second in SR, it can be seen that the proposed tracker obtains an improvement of 4.3%.

Table 1
Evaluation results of the proposed tracker with other latest RGBT trackers on the RGBT234 dataset.

Algorithm	CMPP	SiamCDA	CBPNet	TFNet	Ours
Precision Score	75.1	76.0	79.4	80.6	**82.5**
Success Score	49.1	56.9	54.1	56.0	**61.6**

Table 2
Evaluation results and running efficiency of different trackers on the RGBT234 dataset.

Algorithm	SOWP	MDNet+RGBT	MANet	JMMAC	Ours
Precision Score	64.2	72.2	77.7	79.0	**82.5**
Success Score	41.1	49.5	53.9	57.3	**61.6**
FPS	3.2	3	2	4	**24.6**

Meanwhile, the proposed obtains an improvement of about 2.9% in PR over the second-best tracker DAFNet. These results indicate that our proposed tracker has strong competitiveness. Moreover, the SOWP+RGBT obtains about 25.4%/4.0% promotion in PR/SR over the SOWP based on RGB information and further confirms that RGBT tracking can be more robust than RGB tracking.

Since the codes of these latest trackers, including CMMP [58], SiamCDA [39], CBPNet [59] and TFNet [60] have not been announced, we can only compare our proposed tracker with them according to the results given by their papers. Table 1 reports the comparative result of the proposed tracker and the newly available RGBT tracker. The tracking results of these latest trackers can achieve 75.1%/49.1%, 76.0%/56.9%, 79.4%/54.1% and 80.6%/56.0%, respectively. It is also worthy to note that our proposed tracker still performs more effectively than those newly reported ones. These experimental results further prove the effectiveness of our proposed tracker.

Furthermore, the tracking speeds of some RGBT trackers are shown in Table 2. It is easily shown that the proposed tracker can run up to 24.6 FPS, much higher than that of the other trackers of SOWP, JMMAC, MDNet+RGBT, and MANet. Thus, the proposed tracker outperforms these RGBT trackers in both accuracy and efficiency.

4.2.2. Attribute-based evaluation

To further prove the robustness of the proposed tracker, we compare the proposed tracker with the state-of-the-art trackers under different attributes. The attributes include no occlusion (NO), background clutter (BC), partial occlusion (PO), low resolution (LR), motion blur (MB), heavy occlusion (HO), camera moving (CM), thermal crossover (TC), deformation (DEF), low illumination

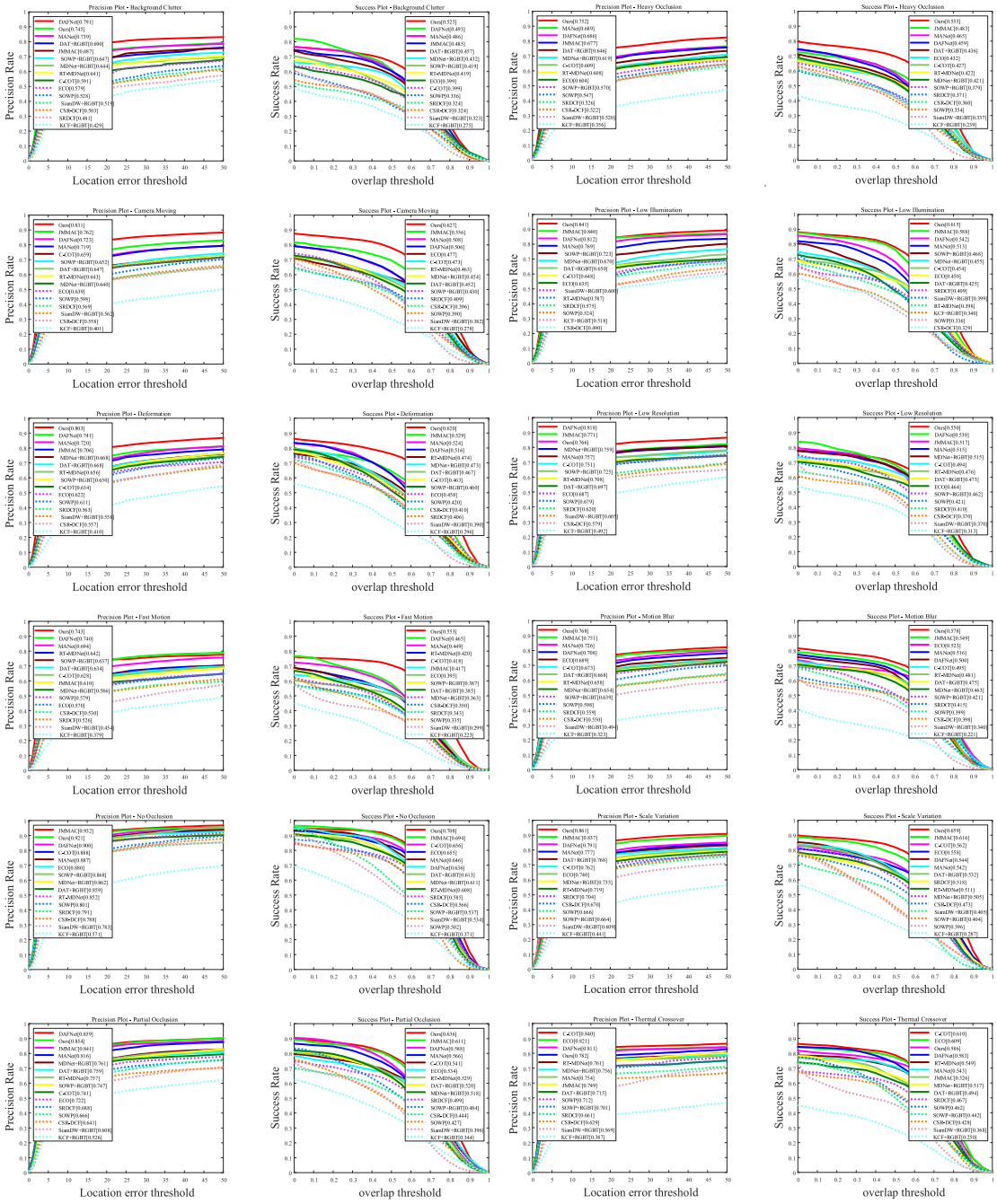

Fig. 9. Attribute-based evaluation plots of different trackers on the RGBT234.

(LI), fast motion (FM) and scale variation (SV). The attribute-based performance comparison results in PR/SR are shown in Fig. 9.

From Fig. 9, we intuitively know that the proposed tracker can deal with the sequences with multiple challenges and significantly perform much better than all other trackers in most attributes. More specifically, most of these trackers can achieve high accuracy facing with no occlusion, while their performance deteriorates rapidly in cases of heavy occlusion. The proposed tracker can still maintain stable robustness which may give the credit to fully using the complementary information of RGB and thermal infrared images. Especially in the cases of CM and DEF, the tracking precision of the proposed tracker can reach 83.1% and 80.3%, respectively, which significantly shows more excellent performance than the second-best tracker in the two cases.

Furthermore, Fig. 10 shows the visualized results in several specific challenging scenarios between the proposed tracker and the others including ECO, MANet, SOWP, etc. It is not difficult to observe the proposed tracker outperforms other trackers in most cases, such as fast motion and partial occlusion. For instance, in the sequence of bikeman, the proposed tracker can accurately locate the target while the compared trackers tend to fail for tracking. This proves that the proposed tracker can achieve outstanding performance in challenging scenarios.

4.3. Evaluation on RGBT210 dataset

To further verify the excellent performance of the proposed tracker, 10 compared trackers are selected on the RGBT210 dataset.

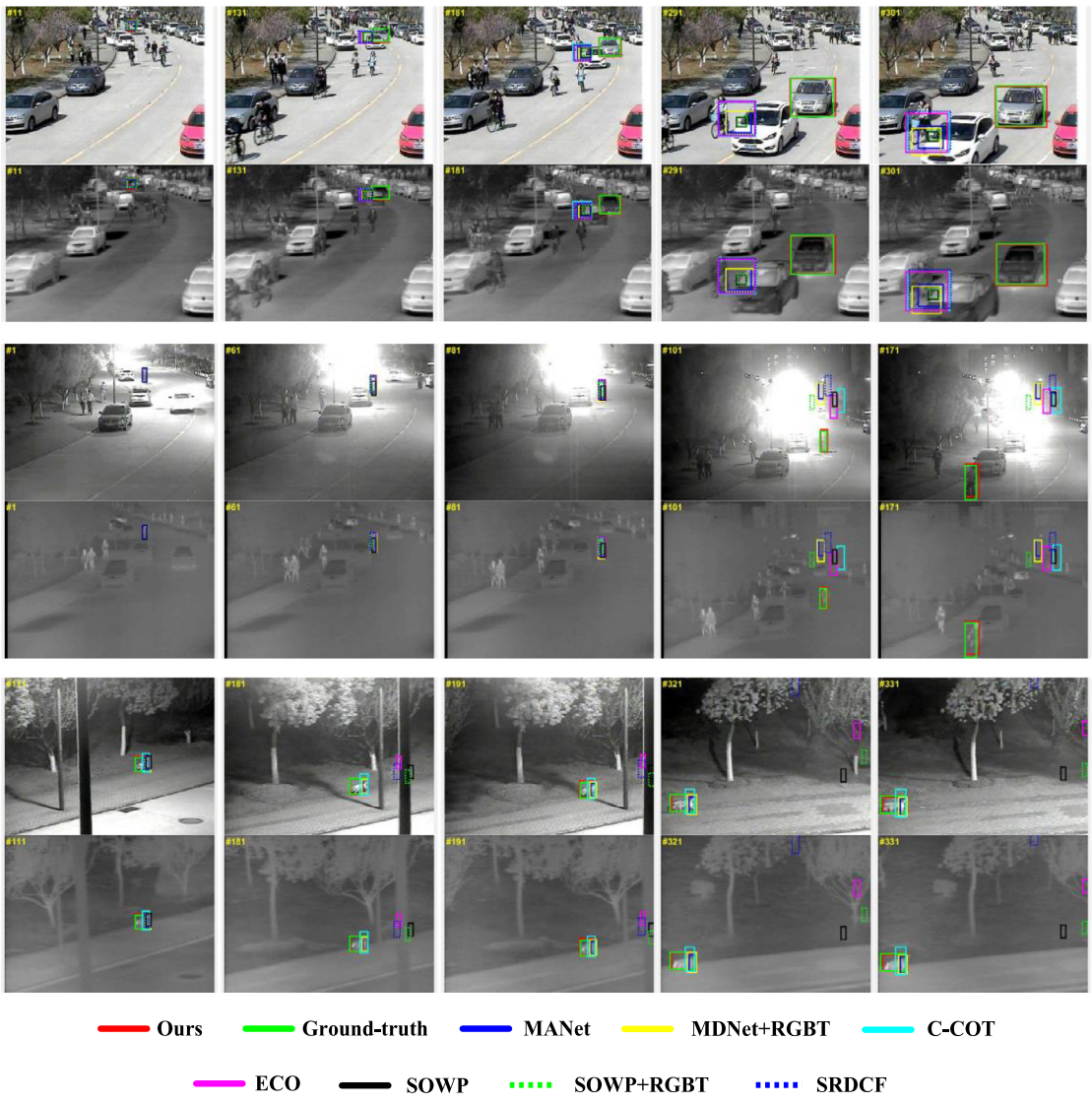

Fig. 10. Visual comparisons of tracking results between our proposed tracker and another three trackers in sequence graycar2, bikeman and dog10 on RGBT234.

They are MANet [57], CCOT [55], MDNet [50], SGT [16], SOWP+RGBT [6], DSST+RGBT [61], MEEM+RGBT [62], KCF+RGBT [56], SOWP [6], SiameseFC [22].

Fig. 11 shows that the proposed tracker obtains 80.6%/59.2% in PR/SR on RGBT210 and outperforms all other trackers. Specifically, compared with the second-best tracker, i.e., MANet, the proposed tracker obtains 5.3%/7.5% promotion in PR/SR. Moreover, the proposed tracker is more robust than the strong trackers like MDNet, SGT, CCOT, and SiameseFC. The top two trackers are all based on deep learning, which also shows that the trackers based on deep learning have gradually become the mainstream.

To further prove the robustness of the proposed tracker, we also compare the proposed tracker with other trackers under each attribute sequence on the RGBT210 and the attribute-based result is shown in Table 3.

As shown in Table 3, the proposed tracker achieves the best result in SR on each attribute. Especially in the challenge of fast motion (FM), the proposed tracker reaches 78.3%/56.2% in PR/SR and obtains 15.0%/16.3% promotion compared to the second-best tracker MANeT. However, the PR score of the proposed tracker is slightly lower than CCOT in thermal crossover (TC). The main reason is that thermal infrared information tends to be unreliable facing thermal crossover. Consequently, the proposed tracker fused the unreliable thermal information tends to be worse than the tracker that only uses the RGB information. This shows that our fusion strategy needs to be further improved. Generally speaking, the proposed tracker is outstanding in almost all attributes, confirming the effectiveness of the proposed tracker.

Furthermore, in Fig. 12, we present the qualitative results on some sequences with different challenging factors between the proposed tracker and other compared trackers including SGT, MDNet, MANet, SiameseFC, etc. As shown in Fig. 12, it can be found that the compared trackers tend to track failure facing the challenging sequences while the proposed tracker can accurately locate the target. This demonstrates that the proposed tracker can obtain outstanding performance in challenging scenarios.

4.4. Ablative study

To provide a thorough analysis of the main components, we divide the proposed tracker into three other versions, including: (1) Baseline, that we extend the Transt-Tracking [45] into RGBT tracker as the baseline tracker; (2) Baseline+WES, that we utilize the weight optimization strategy to update feature representation; (3) Baseline+CWES, that we combine both the weight optimization strategy and center-loss branch. Table 4 shows the result of different versions.

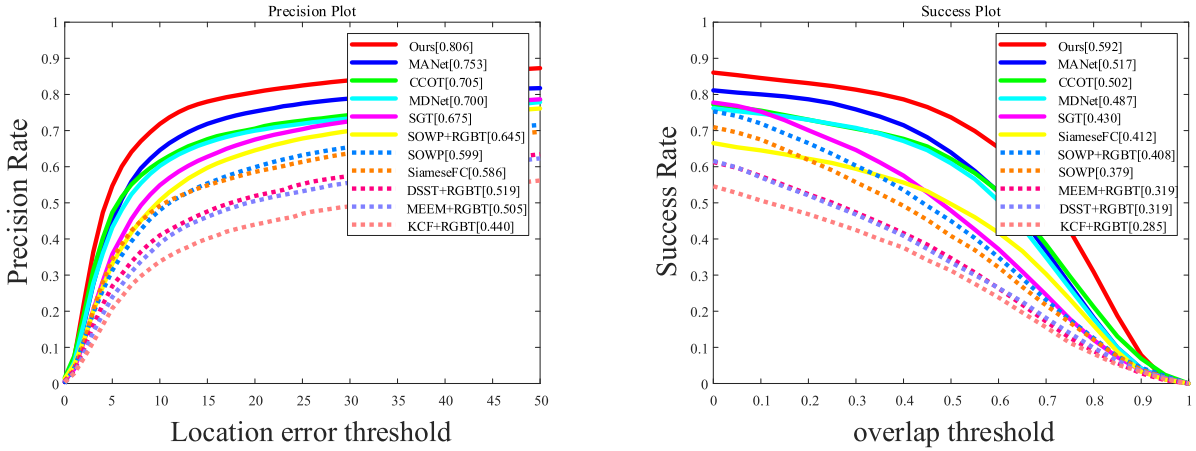

Fig. 11. Evaluation plots of precision and success on the RGBT210 dataset.

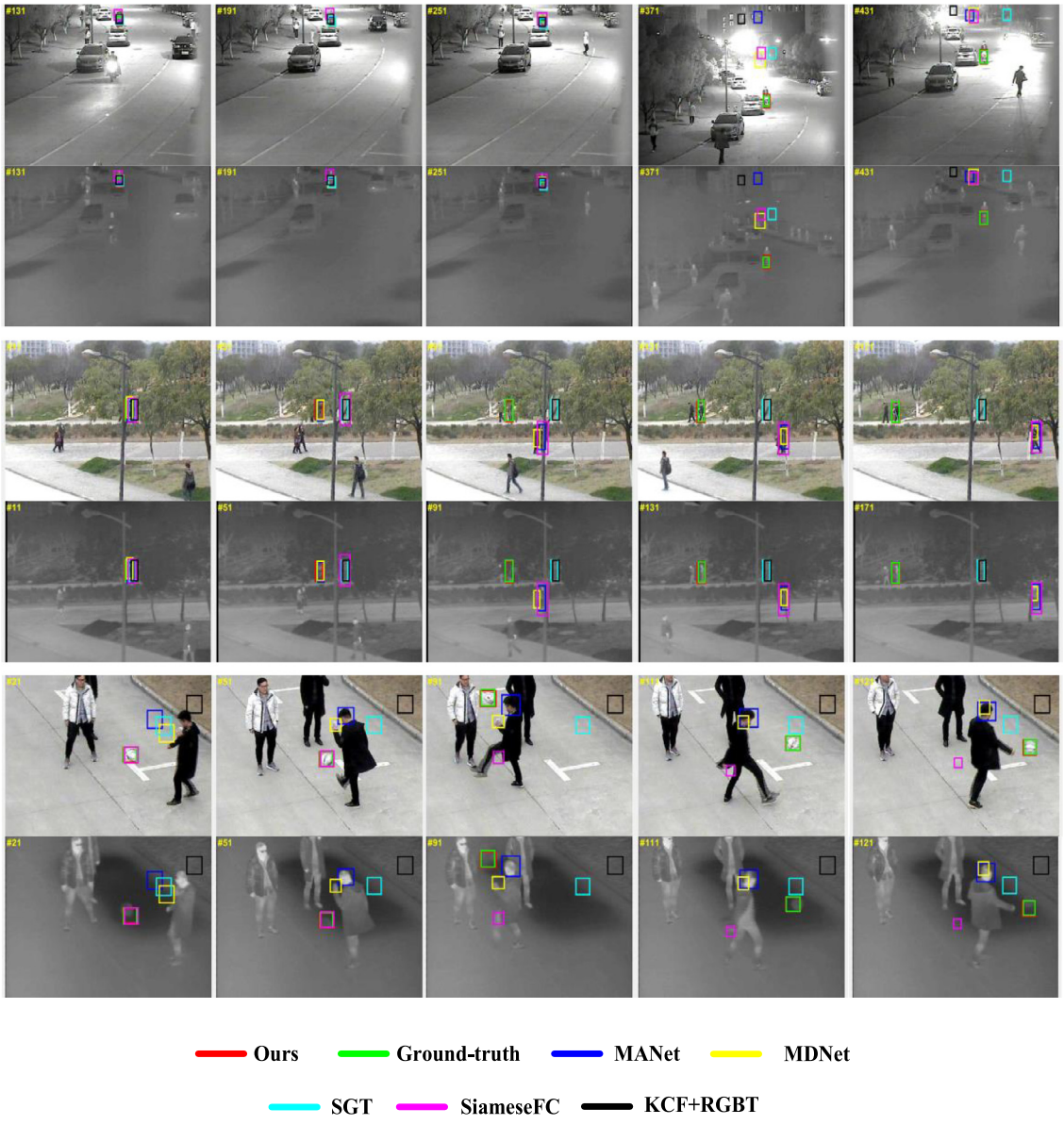

Fig. 12. Visual comparisons of tracking results between our proposed tracker and another three trackers in sequence baby, baketballwaliking and soccer2 on RGBT210.

Table 3
Attribute-based evaluation of the proposed tracker and other state-of-the-art trackers on the RGBT210 dataset. The top three are in red, purple, and blue colors, respectively.

	KCF+RGBT	MEEM+RGBT	DSST+RGBT	SiameseFC	SOWP	SOWP+RGBT	SGT	MDNet	CCOT	MANet	Ours
NO	56.6/36.3	64.7/41.2	68.7/39.0	72.5/53.6	75.0/46.1	79.9/48.4	82.4/50.7	80.2/60.5	85.2/62.5	87.0/62.3	89.2/68.3
PO	49.6/31.6	57.4/35.5	60.7/35.8	62.0/43.6	61.3/39.5	72.5/45.5	75.4/48.3	77.9/53.6	74.0/52.4	82.1/56.2	82.5/60.3
HO	33.0/22.2	37.2/24.2	36.0/25.0	48.9/33.6	52.0/32.8	49.8/32.8	53.1/34.1	57.1/38.6	60.6/42.8	63.2/42.8	75.5/54.5
LI	48.3/30.4	39.2/25.6	56.5/33.5	48.5/34.3	48.3/30.7	69.5/42.7	71.6/44.7	62.6/42.4	66.9/45.0	77.1/50.9	79.4/56.8
LR	42.6/26.2	44.9/23.4	54.6/27.3	40.3/25.0	51.0/29.1	60.8/35.4	65.8/37.5	57.1/36.3	60.0/37.4	67.2/42.3	69.9/46.4
TC	39.0/24.1	58.2/35.6	42.9/25.2	62.8/44.8	70.0/44.9	62.3/39.3	64.9/40.7	72.8/52.2	83.9/58.5	69.3/48.8	80.0/59.4
DEF	40.6/29.5	48.7/33.5	44.9/31.0	54.3/39.8	61.4/41.7	63.1/43.8	65.3/45.9	67.6/48.6	61.1/44.7	69.9/49.2	78.2/60.4
FM	33.3/19.1	43.5/26.8	41.8/24.1	49.0/32.1	56.0/32.3	55.5/32.8	58.0/33.1	59.3/38.5	62.2/41.2	63.3/39.9	78.3/56.2
SV	42.4/27.5	52.8/33.0	55.3/32.6	63.2/45.2	62.8/37.7	63.5/38.6	67.4/41.7	73.3/51.6	76.8/56.7	77.1/54.2	84.5/63.6
MB	29.1/20.7	46.2/31.4	39.0/26.2	50.9/36.7	55.2/38.3	53.7/35.8	58.6/39.6	64.7/46.4	66.6/47.6	64.4/46.0	73.6/54.5
CM	37.5/26.0	48.7/31.9	43.1/29.1	51.9/37.2	55.8/36.9	56.5/38.1	59.0/40.7	62.7/44.8	61.9/44.4	67.3/47.5	79.8/59.7
BC	41.0/25.6	40.5/23.4	48.0/28.1	43.3/29.2	47.2/28.6	58.1/35.9	58.6/35.5	55.7/36.4	54.3/35.6	69.2/44.7	72.6/49.8
ALL	44.0/28.5	50.5/31.9	51.9/31.9	58.6/41.2	59.9/37.9	64.5/40.8	67.5/43.0	70.0/48.7	70.5/50.2	75.3/51.7	80.6/59.2

Table 4
PR(%) and SR(%) scores of the proposed tracker with its variants on RGBT234 and RGBT210 datasets.

Algorithm	Baseline	Baseline+WES	Baseline+CWES(Ours)
RGBT234	77.3/57.2	80.6/60.2	82.5/61.6
RGBT210	75.5/54.8	77.9/57.2	80.6/59.2

Table 5
The running efficiency of different fusion stages on the RGBT234 dataset.

	Middle Fusion	Late Fusion	Ours
FPS	19.4	18.6	**24.6**

As shown in Table 4, the proposed tracker obtains more outstanding performance than other competing versions on RGBT234 and RGBT210 datasets, which demonstrates the robustness of the proposed tracker. More specially, on the RGBT234 dataset, the Baseline+WES obtains the improvement of 3.3% and 3.0% over the Baseline in PR/SR, which fully validates the effectiveness of the proposed weight optimization strategy. Moreover, we find that the Baseline+CWES can obtain more outstanding performance than Baseline+WES, proving the effectiveness of the center-loss branch. Also, Table 4 presents the evaluation results of each component on the RGBT210 dataset. It is easy to observe that the Baseline+CWES obtains about 2.9% and 4.5% promotion in PR/SR over the Baseline+WES. Meanwhile, the Baseline+WES outperforms the Baseline. The above results all fully prove the effectiveness of the weight optimization strategy and center-loss branch.

To prove that the designed shallow convolutional network in early fusion stage can improve the tracking efficiency, we compare the proposed tracker and the variants of the proposed tracker which use middle fusion and late fusion, respectively. Table 5 reports the experimental results of tracking efficiency of different fusion stages trackers. It is easily shown that the proposed tracker can run up to 24.6 FPS, much higher than middle fusion and late fusion. This proves the effectiveness of designed shallow convolutional network in the early fusion stage. Consequently, we can conclude that all the designed modules in the proposed tracker are contributed to improving the performance.

4.5. Evaluation on GTOT dataset

Fig. 13 presents the compared results of the proposed tracker with 9 state-of-the-art trackers including MANet [57], JMMAC [51], CCOT [55], DAT+RGBT [49], MDNet+RGBT [50], RT-MDNet [48], SRDCF [52], SiamDW+RGBT [28] and ECO [53] on the GTOT dataset.

As shown in Fig. 13, the proposed tracker achieves the best performance on the GTOT dataset on both evaluation metrics compared with all other recent trackers. More specifically, the proposed tracker can be up to 91.1%/75.3% and achieves about 0.9%/2.1% improvement in PR/SR over the most competitive tracker JMMAC. Besides, compared with the other recent strong trackers, i.e., MANet and DAFNet, the proposed tracker outperforms them with 1.7%/2.9% and 2.0%/4.1% in PR/SR, respectively. Since the result of the latest Siamese tracker, i.e., SiamCDA has not been announced, we compare the proposed tracker with it, which achieves 87.7%/73.2% in PR/SR, given by Zhang et al. [39]. It is easy to find that the proposed tracker still obtains 3.4%/2.1% promotion over SiamCDA. These experimental results in both precision scores and success scores prove the effectiveness of the proposed tracker.

4.6. Evaluation on LasHeR dataset

The proposed tracker is further compared in precision and success with 12 state-of-the-art trackers including DMCNet [63], MaCNet [64], MANet++ [36], MANet [57], CAT [65], DAFNet [47], mfDiMP [66], FANet [67], DAPNet [68], SGT++ [15], CMR [69] and SGT [16] on the LasHeR testing set, which is the largest RGBT tracking dataset at present.

Fig. 14 reports the compared results of these trackers on the LasHeR test dataset. It can be found that the proposed tracker attains 78.0%/62.6% and outperforms all the compared trackers. Compared with the second-best tracker DMCNet, the proposed tracker obtains an improvement of about 29.0%/27.1%. In addition, the proposed tracker also achieves better tracking results compared with MANet and DAFNet in PR/SR, which are all popular and strong trackers proposed in recent years. These experimental results fully demonstrate the advantages of the proposed tracker.

To make the comparison experiments fair enough and further prove the effectiveness of the proposed tracker, we compare the proposed tracker with the two most representative RGBT trackers, MANet and mfDiMP which are retrained on the LasHeR training set. Table 6 reports the experimental results of the proposed tracker and retrained RGBT trackers on the LasHeR testing set. It is easy to see that the proposed tracker still shows the best excellent performance on the LasHeR testing set compared with the retrained MANet and mfDiMP. More specifically, the proposed tracker can be up to 78.0%/62.6% in PR/SR and has a significant improvement of 17.3%/16.5% and 23.8%/25.8% compared with MANet and mfDiMP. These comparison experiments also fully confirm the robustness and effectiveness of the proposed tracker.

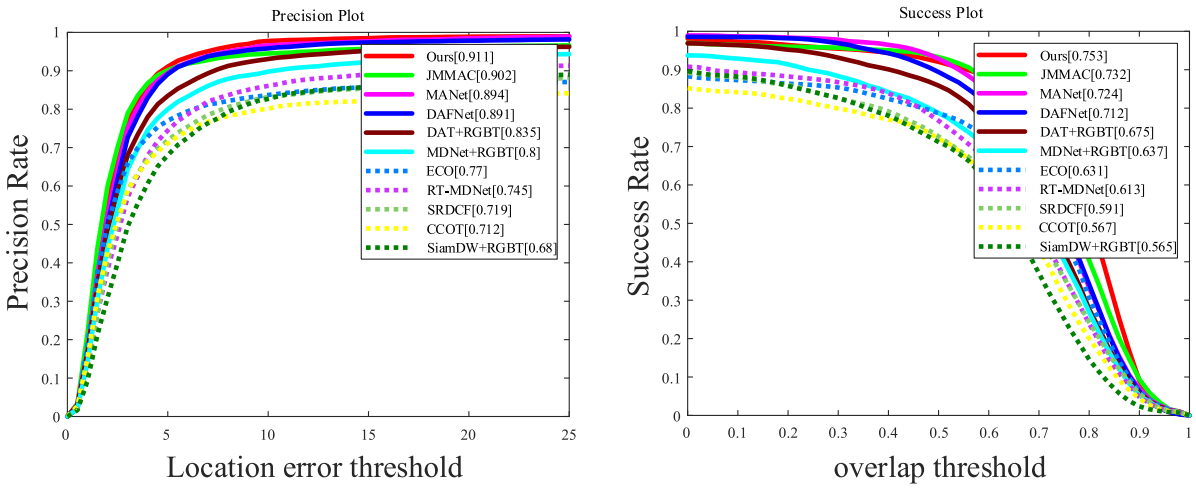

Fig. 13. Evaluation plots of precision and success on the GTOT dataset.

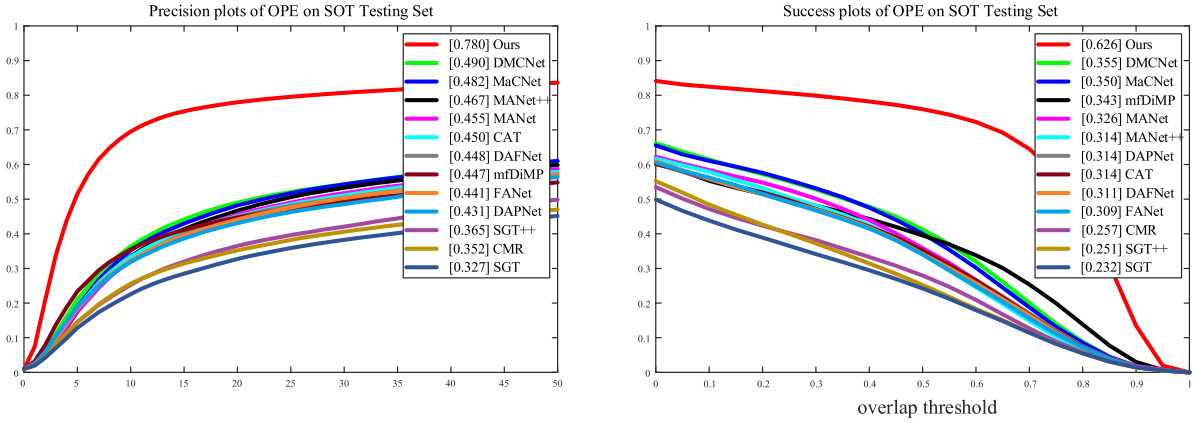

Fig. 14. Evaluation plots of precision and success on the LasHeR test dataset.

Table 6
PR(%) and SR(%) scores of the proposed tracker with retrained mfDiMP and MANet on the LasHeR test dataset.

Algorithm	MANet	mfDiMP	Ours
Precision Score	60.7	54.2	**78.0**
Success Score	46.1	36.8	**62.6**

5. Conclusion

In this paper, we propose a strong cross-modal model based on transformer for robust RGBT tracking. A simple dual-flow convolutional network is first designed to extract and fuse dual-modal features. This effectively reduces the computational complexity in the fusion process. Then, a modal weight allocation strategy is designed to enhance the feature representation. To decrease the loss of semantic information and enhance the connection between long-distance information, a feature fusion network based on transformer is utilized. This helps focus on more discriminating information. Besides, a center-loss branch is adopted in the designed classification and regression subnetwork to more accurately predict the localization of the target. Extensive experiments on RGBT234, RGBT210, GTOT and LasHeR datasets, demonstrate the effectiveness of the proposed tracker against the state-of-the-art trackers.

CRediT authorship contribution statement

Mingzheng Feng: Conceptualization, Methodology, Software, Data curation, Visualization, Writing – original draft. **Jianbo Su:** Supervision, Project administration, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the key project of the National Natural Science Foundation of China under Grant 61533012 and Grant 91748120.

References

[1] L. Yang, C. Kong, X. Chang, S. Zhao, Y. Cao, S. Zhang, Correlation filters with adaptive convolution response fusion for object tracking, Knowl.-Based Syst. 228 (2021) 107314.
[2] N. Wang, W. Zhou, J. Wang, H. Li, Transformer meets tracker: Exploiting temporal context for robust visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1571–1580.

[3] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6668–6677.
[4] K. Dai, D. Wang, H. Lu, C. Sun, J. Li, Visual tracking via adaptive spatially-regularized correlation filters, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4670–4679.
[5] F. Li, C. Tian, W. Zuo, L. Zhang, M.-H. Yang, Learning spatial–temporal regularized correlation filters for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4904–4913.
[6] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, C.-S. Kim, Sowp: Spatially ordered and weighted patch descriptor for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3011–3019.
[7] C. Li, X. Wu, N. Zhao, X. Cao, J. Tang, Fusing two-stream convolutional neural networks for RGB-T object tracking, Neurocomputing 281 (2018) 78–85.
[8] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, H. Wang, Learning deep multi-level similarity for thermal infrared object tracking, IEEE Trans. Multimed. 23 (2020) 2114–2126.
[9] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, W. Liu, Y. Liang, Multi-task driven feature models for thermal infrared tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11604–11611.
[10] H. Liu, F. Sun, Fusion tracking in color and infrared images using joint sparse representation, Sci. China Inf. Sci. 55 (2012) 590–599.
[11] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, L. Lin, Learning collaborative sparse representation for grayscale-thermal tracking, IEEE Trans. Image Process. 25 (2016) 5743–5756.
[12] S. Zhai, P. Shao, X. Liang, X. Wang, Fast RGB-T tracking via cross-modal correlation filters, Neurocomputing 334 (2019) 172–181.
[13] M. Feng, K. Song, Y. Wang, J. Liu, Y. Yan, Learning discriminative update adaptive spatial–temporal regularized correlation filter for RGB-T tracking, J. Vis. Commun. Image Represent. 72 (2020) 102881.
[14] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, G. Xiao, SiamFT: An RGB-infrared fusion tracking method via fully convolutional siamese networks, IEEE Access 7 (2019) 122122–122133.
[15] C. Li, X. Liang, Y. Lu, N. Zhao, J. Tang, RGB-T object tracking: Benchmark and baseline, Pattern Recognit. 96 (2019) 106977.
[16] C. Li, N. Zhao, Y. Lu, C. Zhu, J. Tang, Weighted sparse representation regularized graph learning for RGB-T object tracking, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1856–1864.
[17] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, D. Sun, LasHeR: A large-scale high-diversity benchmark for RGBT tracking, IEEE Trans. Image Process. 31 (2022) 392–404.
[18] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Atom: Accurate tracking by overlap maximization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4660–4669.
[19] G. Bhat, M. Danelljan, L.V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6182–6191.
[20] Q. Liu, X. Lu, Z. He, C. Zhang, W.-S. Chen, Deep convolutional neural networks for thermal infrared object tracking, Knowl.-Based Syst. 134 (2017) 189–198.
[21] Q. Liu, D. Yuan, N. Fan, P. Gao, X. Li, Z. He, Learning dual-level deep representation for thermal infrared tracking, IEEE Trans. Multimed. (2022) http://dx.doi.org/10.1109/TMM.2022.3140929.
[22] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H. Torr, Fully-convolutional siamese networks for object tracking, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 850–865.
[23] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H. Torr, End-to-end representation learning for correlation filter based tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2805–2813.
[24] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, S. Maybank, Learning attentions: residual attentional siamese network for high performance online visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4854–4863.
[25] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[26] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 101–117.
[27] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, Siamrpn++: Evolution of siamese visual tracking with very deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4282–4291.
[28] Z. Zhang, H. Peng, Deeper and wider siamese networks for real-time visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4591–4600.
[29] S. Cheng, B. Zhong, G. Li, X. Liu, Z. Tang, X. Li, J. Wang, Learning to filter: Siamese relation network for robust tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4421–4431.
[30] Y. Wu, E. Blasch, G. Chen, L. Bai, H. Ling, Multiple source data fusion via sparse representation for robust visual tracking, in: Proceedings of the 14th International Conference on Information Fusion, 2011, pp. 1–8.
[31] X. Yun, Y. Sun, X. Yang, N. Lu, Discriminative fusion correlation learning for visible and infrared tracking, Math. Probl. Eng. 2019 (2019) 1–11.
[32] C. Luo, B. Sun, K. Yang, T. Lu, W.-C. Yeh, Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme, Infrared Phys. Technol. 99 (2019) 265–276.
[33] Q. Xu, Y. Kuai, J. Yang, X. Deng, Enhanced real-time RGB-T tracking by complementary learners, J. Circuits Syst. Comput. 30 (2021) 2150307.
[34] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, J. Han, RGB-T salient object detection via fusing multi-level CNN features, IEEE Trans. Image Process. 29 (2019) 3321–3335.
[35] Q. Zhang, T. Xiao, N. Huang, D. Zhang, J. Han, Revisiting feature fusion for RGB-dt salient object detection, IEEE Trans. Circuits Syst. Video Technol. 31 (2020) 1804–1818.
[36] A. Lu, C. Li, Y. Yan, J. Tang, B. Luo, RGBT tracking via multi-adapter network with hierarchical divergence loss, IEEE Trans. Image Process. 30 (2021) 5613–5625.
[37] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, F. Wu, MFGNet: Dynamic modality-aware filter generation for RGB-T tracking, 2021, arXiv preprint arXiv:210710433.
[38] X. Zhang, P. Ye, S. Peng, J. Liu, G. Xiao, DSiamMFT: An RGB-T fusion tracking method via dynamic siamese networks using multi-layer feature fusion, Signal Process., Image Commun. 84 (2020) 115756.
[39] T. Zhang, X. Liu, Q. Zhang, J. Han, SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 1403–1417.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Advances in neural information processing systems, 2017, pp. 5998–6008.
[41] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.
[42] M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, H. Dong, End-to-end object detection with adaptive clustering transformer, 2020, arXiv preprint arXiv:201109315.
[43] J. Choi, H. Jin Chang, S. Yun, T. Fischer, Y. Demiris, J. Young Choi, Attentional correlation filter network for adaptive visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4807–4816.
[44] Y. Yu, Y. Xiong, W. Huang, M.R. Scott, Deformable siamese attention networks for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6728–6737.
[45] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8126–8135.
[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[47] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, F. Wang, Deep adaptive fusion network for high performance rgbt tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 91–99.
[48] I. Jung, J. Son, M. Baek, B. Han, Real-time mdnet, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 83–98.
[49] S. Pu, Y. Song, C. Ma, H. Zhang, M.-H. Yang, Deep attentive tracking via reciprocative learning, in: Proceedings of the Advances in Neural Information Processing Systems, 2018, pp. 1931–1941.
[50] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4293–4302.
[51] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, X. Yang, Jointly modeling motion and appearance cues for robust RGB-T tracking, IEEE Trans. Image Process. 30 (2021) 3335–3347.
[52] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4310–4318.
[53] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 6638–6646.
[54] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, M. Kristan, Discriminative correlation filter with channel and spatial reliability, Int. J. Comput. Vis. 126 (2018) 671–688.

[55] M. Danelljan, A. Robinson, F.S. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 472–488.
[56] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 583–596.
[57] C. Li, A. Lu, A. Zheng, Z. Tu, J. Tang, Multi-adapter RGBT tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2262–2270.
[58] R. Yang, X. Wang, C. Li, J. Hu, J. Tang, RGBT tracking via cross-modality message passing, Neurocomputing 462 (2021) 365–375.
[59] Q. Xu, Y. Mei, J. Liu, C. Li, Multimodal cross-layer bilinear pooling for RGBT tracking, IEEE Trans. Multimed. 24 (2022) 567–580.
[60] Y. Zhu, C. Li, J. Tang, B. Luo, L. Wang, RGBT tracking by trident fusion network, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 579–592.
[61] M. Danelljan, G. Häger, F. Khan, M. Felsberg, Accurate scale estimation for robust visual tracking, in: British Machine Vision Conference, Bmva Press, Nottingham, 2014, pp. 1–5, 2014.
[62] J. Zhang, S. Ma, S. Sclaroff, MEEM: robust tracking via multiple experts using entropy minimization, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 188–203.
[63] A. Lu, C. Qian, C. Li, J. Tang, L. Wang, Duality-gated mutual condition network for RGBT tracking, 2020, arXiv preprint arXiv:201107188.
[64] H. Zhang, L. Zhang, L. Zhuo, J. Zhang, Object tracking in RGB-T videos using modal-aware attention network and competitive learning, Sensors 20 (2020) 393.
[65] C. Li, L. Liu, A. Lu, Q. Ji, J. Tang, Challenge-aware RGBT tracking, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 222–237.
[66] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, F. Shahbaz Khan, Multi-modal fusion for end-to-end rgb-t tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 2252–2261.
[67] Y. Zhu, C. Li, J. Tang, B. Luo, Quality-aware feature aggregation network for robust RGBT tracking, IEEE Trans. Intell. Veh. 6 (2020) 121–130.
[68] Y. Zhu, C. Li, B. Luo, J. Tang, X. Wang, Dense feature aggregation and pruning for rgbt tracking, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 465–472.
[69] C. Li, C. Zhu, Y. Huang, J. Tang, L. Wang, Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 808–823.