International Journal of Machine Learning and Cybernetics (2023) 14:3281–3293
https://doi.org/10.1007/s13042-023-01833-6

ORIGINAL ARTICLE

Siamese infrared and visible light fusion network for RGB-T tracking

Jingchao Peng[1] · Haitao Zhao[1] · Zhengwei Hu[1] · Yi Zhuang[1] · Bofan Wang[1]

Received: 9 August 2021 / Accepted: 6 April 2023 / Published online: 24 April 2023

Abstract
Due to the different photosensitive properties of infrared and visible light, infrared and visible light images have individual features. However, since the registered RGB-T image pairs shot in the same scene, they also contain common features. This paper proposes a Siamese infrared and visible light fusion Network (SiamIVFN) for RBG-T image-based tracking. SiamIVFN contains two main subnetworks: a complementary-feature-fusion network (CFFN) and a contribution-aggregation network (CAN). CFFN utilizes a two-stream multilayer convolutional structure that separately extracts individual features, and filters in each layer are partially coupled to extract common features. CFFN is a feature-level fusion network, which can cope with the misalignment of the RGB-T image pairs. Through adaptively calculating the contributions of infrared and visible light features obtained from CFFN, CAN makes the tracker robust under various light conditions. Experiments show that compared to state-of-the-art techniques, SiamIVFN improves the PR/SR score with 1.5%/8.8% on RGBT234 and 2.1%/6.9% on GTOT. The tracking speed of SiamIVFN is 147.6FPS, the current fastest RGB-T fusion tracker. The source codes are available at https://github.com/PengJingchao/SiamIVFN.

Keywords Object tracking · Deep learning · Fusion tracking · Siamese network

1 Introduction

Object tracking is an essential task in computer vision [1–3]. In the past decades, deep convolutional networks have been successfully applied in different fields, especially in object tracking [4–7]. Visible light images have rich texture information and high contrast, which is useful for object tracking. However, in weak light conditions such as cloudy nights or low visibility conditions such as aerosols, visible-light-images-based object tracking may be difficult to function. Unlike visual light images, thermal infrared images, which mainly record the thermal radiation of objects, are stable under drastic changes in weak light or low visibility conditions [8, 9]. Infrared radiation can penetrate rain, fog, and snow. Nevertheless, infrared images lack texture information rich in visible light images and have low contrast. Due to the complementarity between infrared and visible light images, object tracking based on the fusion of infrared and visible light images has attracted more and more attention [10–13].

The existing fusion tracking based on infrared and visible light images (or so-called RGB-T fusion tracking) methods focuses on supplementing thermal information to assist visible-light-image-based tracking [14–16]. They aim to compensate for the visible light image in deteriorated light conditions. RGB-T fusion can be divided into pixel-level fusion, feature-level fusion, and decision-level fusion [17].

Pixel-level fusion fuses the rigorous registered image pairs pixel by pixel and then performs object tracking based on the merged images [17]. Pixel-level fusion is sensitive to noise and has a high demand for image registration [18]. Decision-level fusion performs tracking tasks separately on RGB and thermal images and then aggregates two different tracking results (such as the position and the size of the tracking object) to obtain the final tracking result [19]. That is, separate images are processed individually and fed into the fusion algorithm. Unlike pixel-level fusion, decision-level fusion does not require obtaining individual pixel values at the same locations by interpolation. However, decision-level fusion pays little attention to the feature complementarity between infrared and visible light images,

✉ Haitao Zhao
haitaozhao@ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Meilong Road 130, Shanghai 200237, China

Fig. 1 Superimposed image of infrared and visible light image pair. The weight of the visible light image is 0.6, and the weight of the infrared image is 0.4. The object in the yellow box with rich texture, details, and color is from the visible light image. The object in the red box, which the silhouette can roughly distinguish, is from the infrared image

leading to unreliable tracking effects that rely on a single pattern [18].

Feature-level fusion is to extract and fuse features of RGB-T image pairs and then utilize the fused features for tracking [20–22]. In this way, the tracking result may not rely too much on a single pattern tracker or original strict registered image pairs [17, 18]. Although a spatial registration step is still necessary for feature extraction and fusion, feature-level fusion allows for explicitly handling localization uncertainty, for instance, due to the misalignment of the image pairs.

To visually demonstrate the characteristics of the RGB-T pairs, we linearly superimpose the infrared and visible light images, as shown in Fig. 1. On the one hand, the image pairs contain common features because two images are shot simultaneously at the same place. On the other hand, since cameras capture two images with different sensor types, the infrared image and the visible light image have individual features: the visible light image is a high-resolution color image with rich textures and details, while the infrared image is monochrome, low contrast, and lack textures. Besides, because of the different clock frequencies of the different photosensitive chips, even if the two cameras are registered in advance, the position of the same object in the two images is not necessarily the same. Unregistered cases require that the fusion tracker have a certain ability of misalignment prevention during the fusion process.

Based on the above analysis, a feature-level fusion network, Siamese Infrared and Visible Light Fusion Network (SiamIVFN) is proposed to track an object in RGB-T image pairs. The feature fusion part of SiamIVFN is composed of two subnetworks: a complementary-feature-fusion network (CFFN) and a contribution-aggregation network (CAN). CFFN uses a two-stream convolution structure to extract and fuse the features of infrared and visible light images. In each layer of the two-stream convolution, a coupled filter is designed to extract the common features from the image pairs. Considering that the similarity of the features extracted from the shallow layers in the RGB-T image pairs is different from the features extracted from the deep layers, we gradually increase the coupling rates. The effectiveness of the coupling rate setting is demonstrated in the experimental part.

Besides, because infrared and visible light images have different contributions to object tracking, further processing of the feature fusion should be considered. CAN uses the self-attention method to adaptively calculate the contribution of infrared and visible light images to different visual conditions. Experiments show that SiamIVFN achieves the best effects of infrared and visible light fusion tracking.

In summary, our contributions are summarized below:

1. CFFN utilizes a two-stream convolutional network with increasing coupling filters to extract the common and individual features. In addition, CFFN has the ability to misalignment prevention.
2. CAN adaptively calculates the contributions of the infrared and visible light features, which can make SiamIVFN robust to various lighting conditions
3. SiamIVFN adopts a Siamese-framework-based fusion tracker for RGB-T fusion tracking, whose structure is straightforward. Therefore SiamIVFN can achieve real-time tracking (tracking speed is 147.6 FPS).

2 Related work

2.1 Visual object tracking

In object tracking, deep learning-based trackers have achieved state-of-the-art performance on multiple public datasets with their powerful representation capabilities. At present, most object trackers based on deep learning have adopted the structure of the Siamese network. SiamFC [4] used the similarity learning method to treat the object tracking problem as a template matching problem. SiamFC is simple and fast. However, since SiamFC uses a multi-scale prediction method, it cannot handle the situation when the size of the object changes drastically. To solve this problem, SiamRPN [5] introduced the region proposal network (RPN). Researchers have improved the Siamese-network-based methods in data preprocessing [23], network structure [24], and multilayer feature fusion [6]. SiamFC++ [7] introduced the concept of anchor-free in the Siamese

network, thus improving the speed and accuracy of the Siamese-framework-based tracker. Although visible light object tracking can achieve good results, it still cannot handle smoke, night, and other bad visual conditions due to the characteristic of RGB sensors.

2.2 RGB-T object tracking

Owing to the penetration ability of infrared sensors, they are often adopted to work with RGB sensors. Therefore, RGB-T fusion tracking has recently attracted more and more attention. SiamFT [21] and DsiamFT [22] used Siamese networks to solve the RGB-T fusion tracking problem. They used different backbones to extract features from infrared and visible light image pairs, then merged them and fed them into the tracking head. Due to the simple structure of the siamese-based methods, the tracking speed of these methods is fast. However, these fusion methods processed the image pairs separately. They did not fully consider the common features of infrared and visible light images, resulting in a lot of feature redundancy and computational burden. Unlike extracting features using different backbones, MANet [20] and CANet [25] shared a part of the same convolution kernels (or so-called coupling filters) to extract common features of infrared and visible light images. However, in designing convolution kernels of different depths, they did not fully consider the features extracted by different depth layers. In this paper, we use different coupling filters in different layers. Besides the consideration of the common feature extraction, the attention mechanism is utilized to extract individual features that reflect the characteristics of the two different sensors. The experimental results of LTDA [26] and CMPP [27] showed that the attention mechanism could largely improve tracking performances.

3 Our method

This section will introduce the proposed SiamIVFN. First, we summarize the overall structure of SiamIVFN and then introduce the structures of CFFN and CAN.

3.1 The architecture

The SiamIVFN network consists of three parts: CFFN, CAN, and tracking head. The structure of SiamIVFN is illustrated in Fig. 2. In the online tracking process, given the infrared and visible light video sequences, the tracker will track the position of the object in each frame. Unlike visible light images, infrared images lack detailed information, such as color and texture. It is necessary to use uncoupled filters to extract individual features from infrared and visible light images separately. Since each infrared and visible image pair simultaneously capture the same scene, they contain common features, such as semantics and contours. Wang [28] argued that coupled filters could extract common features. Li [29] adopted the coupled filters for depth estimation and showed their effectiveness. Inspired by this research work, this paper proposes a complementary-feature-fusion network (CFFN) to extract and fuse the features of infrared and visible light images.

Besides the common features, infrared and visible light images contain individual features for object tracking and may have different contributions to tracking tasks under different light conditions. In degraded light conditions such as fog and night, infrared images contribute more than visible light images for object tracking. While under normal lighting conditions, visible light images are more suitable than infrared images for detecting and tracking an object. Most existing fusion methods regard the contribution of infrared and visible images as the same and often directly concatenate the features extracted from infrared and visible light images. This paper proposes a contribution-aggregation network (CAN), which adaptively calculates the contribution of different features. CAN utilizes the self-attention module [30] to adaptively calculate the contributions of infrared and visible light images according to different light conditions.

3.2 Complementary-feature-fusion network

The details of the CFFN are depicted in Fig. 3. CFFN adopts a two-stream convolution structure. The lower branch represents the convolutional flow of infrared images. The upper branch represents the convolutional flow of visible light images. Unlike other two-stream networks, CFFN sets up filters with different coupling rates in each convolutional layer to learn the common features between infrared and visible images. The overlapping yellow part between the two indicates the coupling part of the two image filters. In this way, infrared and visible light images are mutually auxiliary. The features extracted from the infrared image are supplementary to the stream network designed for the visible light image. In the other stream network for infrared images, features extracted from visible light images are supplementary through partially coupled filters. The uncoupled filters are designed to learn the individual features. The ratio of the number of coupling filters to the number of all filters is called the coupling ratio:

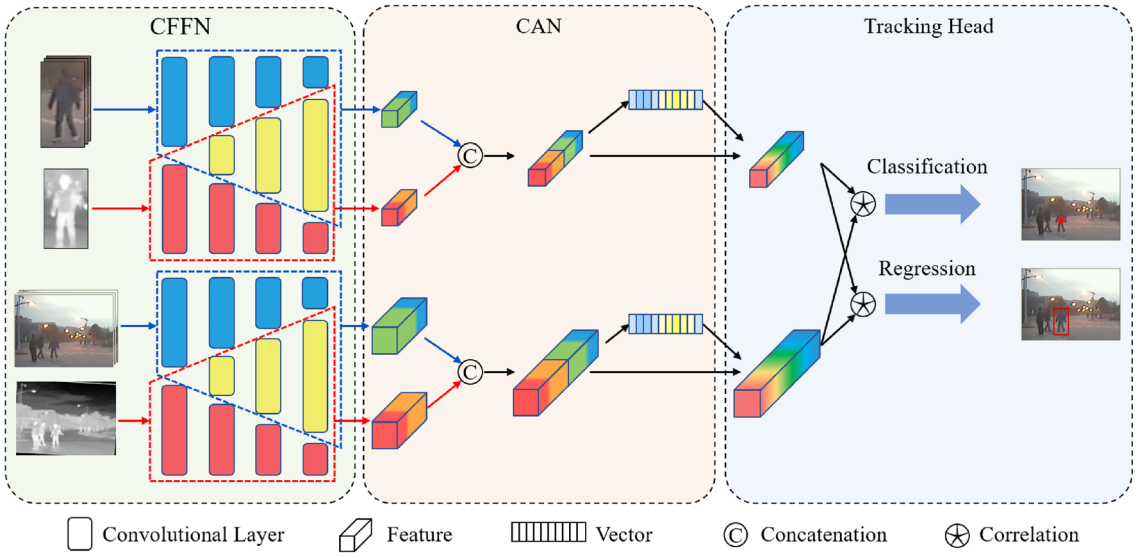


Fig. 2 Illustration of the proposed SiamIVFN framework. The complementary-feature-fusion network (CFFN) is used to extract and fuse the features of RGB-T image pairs. The contribution-aggregation network (CAN) is utilized to calculate the contribution of different features for tracking tasks adaptively. The tracking head is divided into two branches: classification and regression. Please refer to Sect. 4 for more details

$$R_i = \frac{k_i}{n_i}(i = 1, 2, 3, 4), \tag{1}$$

where R_i is the coupling rate of the ith layer, k_i is the number of coupled filters in the ith convolutional layer, and n_i is the number of all filters in the ith convolutional layer. We set the coupling ratio of each convolutional layer: 0.25, 0.5, and 0.75. In Sect. 5, we use the grid search method to demonstrate the effectiveness of the coupling rate. The coupling rate increases as the convolutional layers go deeper. In CFFN, the parameters of the filter are updated through the backpropagation algorithm. In each iteration, the non-coupling filter for infrared and visible light images is updated once, and the coupling filter for infrared and visible light images is updated twice. Therefore, if suppose that we update weights in the infrared stream first and then update weights in the visible light stream, the filter weights are updated as follows:

$$w_{\mathrm{RGB}}^{(i)} = \begin{cases} w_{\mathrm{RGB_{ucoupled}}}^{(i-1)} + l\frac{\partial L}{\partial w_{\mathrm{RGB_{ucoupled}}}^{(i-1)}} \\ w_{\mathrm{coupled}}^{(2i-1)} + l\frac{\partial L}{\partial w_{\mathrm{coupled}}^{(2i-1)}} \end{cases}, \tag{2}$$

$$w_{\mathrm{T}}^{(i)} = \begin{cases} w_{\mathrm{T_{ucoupled}}}^{(i-1)} + l\frac{\partial L}{\partial w_{\mathrm{T_{ucoupled}}}^{(i-1)}} \\ w_{\mathrm{coupled}}^{(2i-2)} + l\frac{\partial L}{\partial w_{\mathrm{coupled}}^{(2i-2)}} \end{cases}, \tag{3}$$

where w is the parameter that needs to be updated, (i) is the number of iterations, l is the learning rate, and L is the loss function. The weights of the coupled filters updated in visible-stream as follows:

$$w_{\mathrm{coupled}}^{(2i-1)} = w_{\mathrm{coupled}}^{(2i-2)} + l\frac{\partial L}{\partial w_{\mathrm{coupled}}^{(2i-2)}}. \tag{4}$$

The weights of the coupled filters updated in infrared-stream as follows:

$$w_{\mathrm{coupled}}^{(2i-2)} = w_{\mathrm{coupled}}^{(2i-3)} + l\frac{\partial L}{\partial w_{\mathrm{coupled}}^{(2i-3)}}. \tag{5}$$

In summary, a two-stream convolutional structure is designed in CFFN. Besides individual features, the two-stream convolutional blocks are able to extract common features by the coupling filters.

3.3 Contribution-aggregation network

After extracting features from infrared and visible light images (using certainly separated backbones), most existing fusion trackers directly concatenate the features and then send them to the tracking head for tracking. However, the different features contribute to object tracking differently, especially under various light conditions. Inspired by SENet [31], this paper proposes CAN to adaptively calculate the contribution of the features, which can be shown in Fig. 4. The difference between CAN and SENet is that CAN adds a

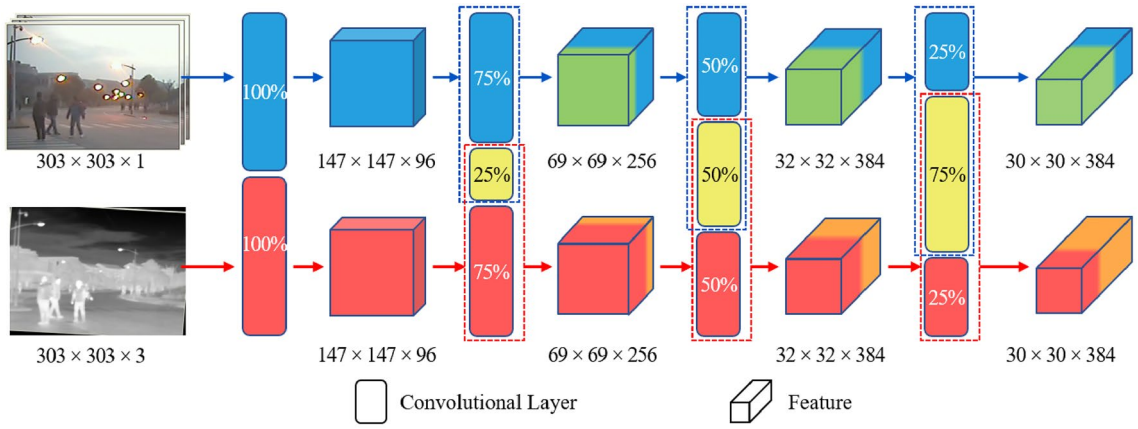


Fig. 3 Illustration of CFFN. CFFN is a two-stream convolutional network, which extracts the features of infrared and visible light images, respectively. The two-stream convolutional network is equipped with coupling filters with different coupling rates

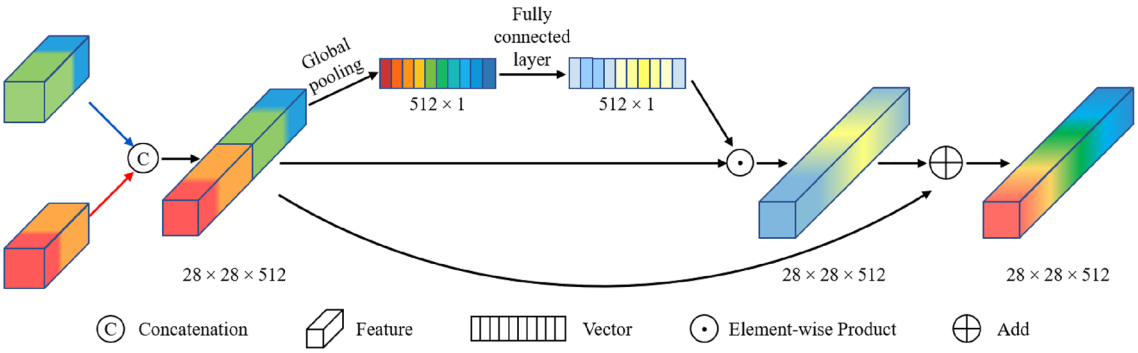


Fig. 4 Illustration of CAN.CAN first concatenates features from RGB and infrared streams. Then, the extracted features are compressed into channel-wise vectors and fed to two fully connected layers. The outputs are multiplied by the original features, finally, added to the original features

step of concatenation. CAN first concatenates features from RGB and infrared streams. Then, CAN utilizes global average pooling to each channel to obtain a global feature g_c as:

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j), \tag{6}$$

where H and W are the height and width of the original feature x_c, respectively. The global feature then passes through two fully connected layers to improve the generalization ability of CAN:

$$h_c = \beta\big(\alpha\big(g_c\big)\big), \tag{7}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are two different fully connected layers. The learned feature vector h_c is multiplied by x_c:

$$y_c = h_c \cdot x_c. \tag{8}$$

Finally, the obtained feature y_c is added to the original feature to calculate the output z_c of CAN:

$$z_c = y_c + x_c. \tag{9}$$

The whole procedure of CAN can be viewed as learning the weight coefficient of each channel through self-attention, which pays more attention to the channels critical for object tracking through end-to-end learning.

4 Implementation details

This section will introduce the training and online tracking process. The tracking head is built based on SiamFC++ [7]. We train and test SiamIVFN on the PyTorch platform with I7-10700K CPU and TITAN RTX GPU.

4.1 Training procedure

4.1.1 Pre-training

We use the GOT10K [32] and LASOT [33] datasets to train our network end-to-end. Since GOT10K and LASOT are both RGB datasets, they do not have infrared images. We use visible light images to generate grayscale images to train coupling filters and non-coupling filters. The optimization algorithm is the stochastic gradient descent method with momentum. The momentum is set to 0.9, and the weight attenuation is set to 0.0001. The learning rate adopts the cosine decay strategy, the initial learning rate is set to 0.08, and the final learning rate is set to 1e–6.

4.1.2 Training

Based on the pre-trained network, we train the entire network using the RGB-T dataset. In the first ten epochs, CFFN is fixed to train CAN and the tracking head. In the second ten epochs, we unfreeze the non-coupling filters for infrared images in CFFN. In the third ten epochs, we unfreeze the coupling filter in CFFN. After the 40th period, we unfreeze the whole CFFN for training. Such gradual training can accelerate the convergence of the network. To improve the discriminative ability of the network, we set the maximum index of a pair of sample frames to 1000 and the ratio of the number of positive sample pairs to the number of negative sample pairs to 0.5. In terms of optimization algorithms, we use Adam to optimize the loss function. The learning rate also uses cosine decay. The initial learning rate is set to 8e-5, and the end learning rate is set to 1e-6.

4.2 Online tracking

In the online tracking process, the template RGB-T image pair and the RGB-T image pair to be searched are fed to CFFN. Then CAN obtain the features of the template and the search area. After the two features are cross-correlated, a score map is computed for classification. According to Xu [7], the direct utilization of the score map for boundary selection might cause performance degradation. In this paper, we adopt a quality estimation branch in addition to the classification branch. The classification branch uses focal loss [34]. The quality estimation branch uses center loss [35]. The 1×1 convolution is used to weight the classification score and quality estimation score to get the overall classification score. In the regression branch, to avoid artificially setting anchor points and thresholds and other manual interventions, we adopt the idea of anchorless to directly predict the four sides from the corresponding position (x, y) in the bounding box. The regression branch uses IOU loss [7].

5 Experiment

5.1 Evaluation dataset and evaluation metrics

We compare SiamIVFN and other tracking methods on two RGB-T tracking benchmark datasets to demonstrate the performances. The GTOT [36] dataset has 15.8K frames, containing 50 RGB-T videos aligned in space and time and seven annotated attributes. The RGBT234 [37] dataset has 234K frames, 234 aligned RGB-T videos, and twelve annotated attributes. Due to the significant differences in the number, quality, and data distribution of GTOT and RGBT234, we divided GTOT and RGBT234 into different training and test sets, respectively. We divide GTOT into five parts, each containing ten videos. When performing experiments on GTOT, we use four parts for training and one for testing. We conduct five separate experiments to ensure that all GTOT datasets are tested. When performing experiments on RGBT234, we divide the dataset into nine parts, each containing 26 videos. Eight parts are utilized for training; one left part is for testing. Nine experiments were performed separately.

The precision rate (PR) and the success rate (SR) in one-pass evaluation (OPE) are used as evaluation indicators. PR refers to the percentage of frames whose distance between the output and ground truth positions is within a threshold. We set the thresholds of GTOT and RGBT234 datasets to 5 and 20, respectively. SR is the proportion of frames whose overlap ratio between the output bounding box and the ground truth bounding box is larger than a threshold. We use the area under the curves (AUC) to calculate the SR score.

5.2 Comparison with state-of-the-art trackers

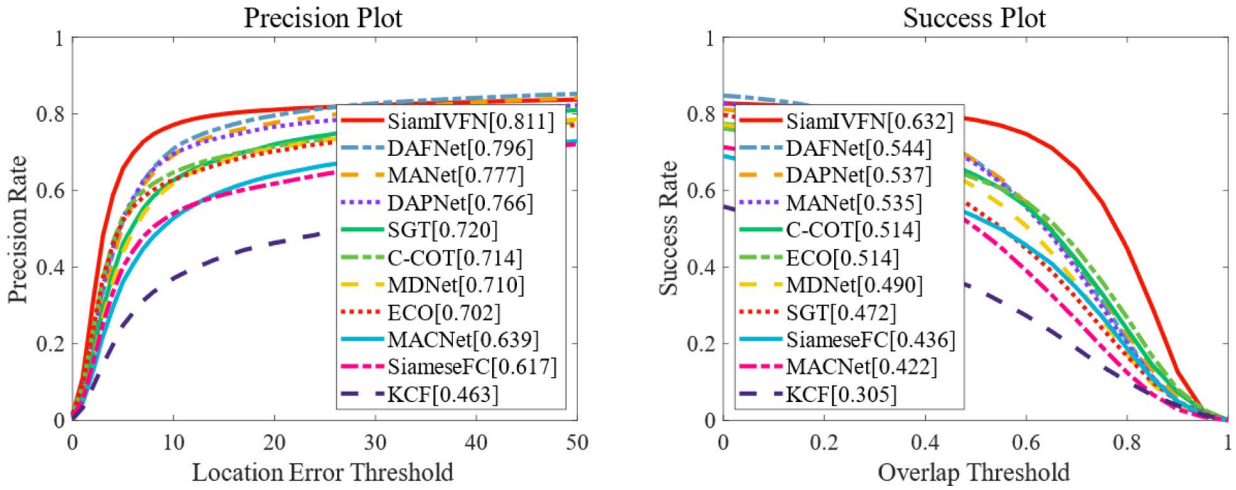
We implemented SiamIVFN on the GTOT and RGBT234 benchmarks and compared them with other state-of-the-art RGB trackers (KCF [38], ECO [39], C-COT [40], MDNet [41], and SiameseFC [4]) and state-of-the-art fusion trackers (SGT [42], MANet [20], MACNet [43], DAPNet [44], and DAFNet [45]). The overall tracking performances are shown in Fig. 5. It can be found that SiamIVFN outperforms other trackers. Specifically, on the RGBT234 dataset, the PR/SR score of SiamIVFN reached 81.1%/63.2%, 1.5%/8.8% higher than the second-best method. As for the GTOT dataset, the PR/SR score of SiamIVFN reached 91.5%/79.3%, 2.1%/6.9% higher than the second-best one. The experimental results demonstrated the effectiveness of the proposed SiamIVFN.

To further show the performances of SiamIVFN, we separately calculated the PR/SR scores of each attribute in the RGBT234 dataset. The specific results are recorded in Table 1. It can be concluded from the table that SiamIVFN

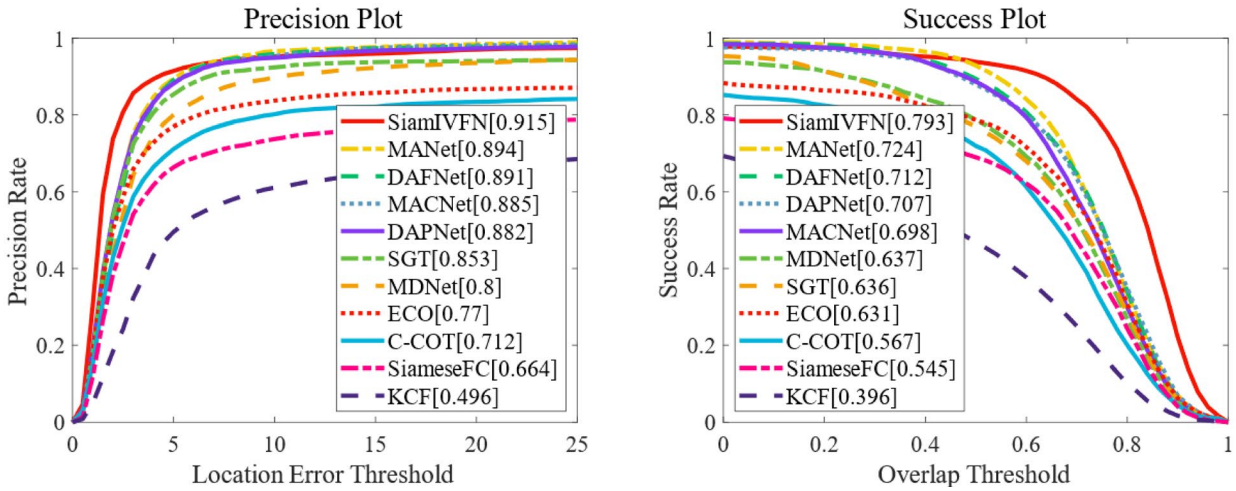
has the highest scores in almost all attributes than other trackers.

Besides the improvement of precision rates, the success rate of the proposed SiamIVFN is much higher than any other tracker (8.8% higher than the second-best), specifically under the challenges of low illumination (LI), low resolution (LR), background clutter (BC), partial occlusion (PO), and heavy occlusion (HO). It means that the proposed subnetworks CFFN and CAN can adaptively extract and fuse the features for the success of object tracking. The two-stream convolutional structure and the channel-wise aggregation are simple and effective for the RGB-T tracking tasks.

In the case of low illumination (LI), relying only on visible light for tracking will lead to poor results. Since SiamIVFN can integrate visible light and infrared images and use infrared image information to supplement tracking, the success rate of SiamIVFN increases by 7.9% compared to the second-best (DAFNet). In the case of background clutter (BC), because of the simple background of the infrared images, SiamIVFN exploits the individual features of the infrared images, and the success rate of SiamIVFN is 8.7% higher than the second-best (DAFNet). In the case of partial occlusion (PO) and heavy occlusion (HO), SiamIVFN can extract the common features and cope with a certain misalignment, thereby increasing tracking robustness. In PO and HO, the success rate of SiamIVFN increases by 6.5% and 12.8% from the second-best (DAFNet for PO, MANet for HO).



(a) Comparison on RGBT234



(b) Comparison on GTOT

Fig. 5 Overall performance compared with state-of-the-art trackers on RGBT234 (**a**) and GTOT (**b**)

Table 1 RGBT234 dataset PR/SR scores based on attributes

PR/SR	KCF	ECO	C-COT	MDNet	SiameseFC	MANet	MACNet	SGT	DAPNet	DAFNet	SiamIVFN
NO	57.1/37.1	88.0/65.5	88.8/65.6	81.2/59.0	76.6/55.5	88.7/64.1	75.4/50.4	87.7/55.5	90.0/64.4	90.0/63.6	86.0/68.5
PO	52.6/34.4	72.2/53.4	74.1/54.1	74.7/50.9	62.8/45.0	81.7/56.2	70.1/46.5	77.9/51.3	82.1/57.4	85.9/58.8	83.0/65.3
HO	35.6/23.9	60.4/43.2	60.9/42.7	63.3/43.2	53.9/36.9	68.8/46.1	53.2/34.6	59.2/39.4	66.0/45.7	68.6/45.9	77.3/58.9
LI	51.8/34.0	63.5/45.0	64.8/45.4	58.9/39.6	58.5/40.2	76.9/51.3	69.7/45.5	70.5/46.2	77.5/53.0	81.2/54.2	81.0/62.1
LR	49.2/31.3	68.7/46.4	73.1/49.4	66.0/44.5	62.6/41.4	76.0/51.1	57.8/35.1	75.1/47.6	75.0/51.0	81.8/53.8	73.9/55.4
TC	38.7/25.0	82.1/60.9	84.0/61.0	74.8/53.0	61.2/43.2	75.4/53.8	48.4/31.9	76.0/47.0	76.8/54.3	81.1/58.3	64.9/48.3
DEF	41.0/29.6	62.2/45.8	63.2/46.3	66.4/46.8	58.1/42.6	72.0/52.0	61.6/42.5	68.5/47.4	71.7/51.8	74.1/51.6	79.6/63.0
FM	37.9/22.3	57.0/39.5	62.8/41.8	63.2/39.3	57.5/37.6	69.5/44.5	55.7/33.3	67.7/40.2	67.0/44.3	74.0/46.5	67.0/48.3
SV	44.1/28.7	74.0/55.8	76.2/56.2	73.9/51.9	63.0/45.2	77.7/53.9	65.0/42.8	69.2/43.4	78.0/54.2	79.1/54.4	82.9/65.3
MB	32.3/22.1	68.9/52.3	67.3/49.5	62.4/44.2	56.4/40.8	72.6/51.3	46.8/32.2	64.7/43.6	65.3/46.7	70.8/50.0	63.3/49.7
CM	40.1/27.8	63.9/47.7	65.9/47.3	61.3/43.3	58.0/41.8	71.9/50.3	55.5/37.9	66.7/45.2	66.8/47.4	72.3/50.6	70.6/54.7
BC	42.9/27.5	57.9/39.9	59.1/39.9	62.5/41.8	50.2/33.7	73.8/48.0	57.3/34.7	65.8/41.8	71.7/48.4	79.1/49.3	76.9/58.0
ALL	46.3/30.5	70.2/51.4	71.4/51.4	71.0/49.0	61.7/43.6	77.7/53.5	63.9/42.2	72.0/47.2	76.6/53.7	79.6/54.4	81.1/63.2

The best, second-best, and third-best PR (SR) are shown in red, blue, and yellow

5.3 Ablation Study

In this subsection, we compared the tracking performances of SiamIVFN (RGB), SiamIVFN (T), and SiamIVFN. SiamIVFN (RGB) and SiamIVFN (T) indicate that SiamIVFN relies solely on visible light and infrared images for tracking, respectively. SiamIVFN (RGB) refers to replacing the infrared image part with the visible light image. SiamIVFN (T) refers to replacing the visible light image part with the infrared image. The tracking performance is shown in Fig. 6a. The experimental results show that SiamIVFN fusion tracking is significantly better than relying solely on infrared images (12.3%/12.6) or visible light images for tracking (9.0%/14.2%).

To show the performance of the two subnetworks, CFFN and CAN, we remove CFFN and CAN from SiamIVFN [denoted by SiamIVFN (No-CFFN) and SiamIVFN (No-CAN)]. Comparative experiments are performed on the RGBT234 dataset. SiamFC++(RGBT) is the baseline. In SiamFC++(RGBT), the infrared image is directly used as the fourth channel, concatenated on the RGB image, and then tracked by SiamFC++. The results are shown in Fig. 6b, which show that:

1. Comparing the performance of SiamIVFN and SiamIVFN (No-CAN), the PR/SR score with CAN improves by 1.6%/3.0%.
2. Comparing the performance of SiamIVFN and SiamIVFN (No-CFFN), the PR/SR score with CFFN improves by 9.3%/8.7%.

The coupling rate of different layers in the CFFN is a hyper-parameter of SiamIVFN. We arrange the coupling rates 0.25, 0.5, 0.75 in separate layers and then compare networks with different coupling rates. The tracking performance under RGBT234 is shown in Table 2. From Table 2, we can find that the greater the coupling rate in deep layers is, the better the tracking performance is. When the coupling ratio is 0.25, 0.5, 0.75, the tracker can obtain the best performance. The features extracted by the shallow layers are individual features such as color and texture. These individual features between infrared and visible light images are quite different, so the appropriate coupling rate is small. On the contrary, the common features such as the contour extracted by the deep network between infrared and visible light images, are relatively similar, so the appropriate coupling rate is larger.

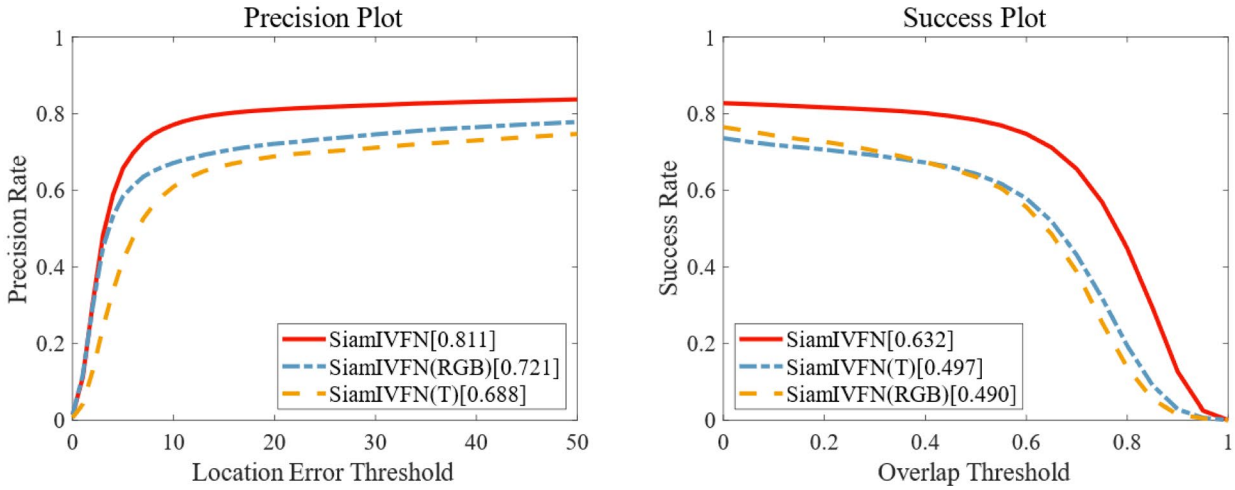
5.4 Qualitative performances

To visually show the tracking performances of SiamIVFN, we took four sequences for comparison. Figure 7 shows the bounding box of SiamIVFN and other trackers (MANet, C-COT, SiamFC, and SiamFC++). To show the bounding boxes in one image, we linearly superimpose the infrared and visible light images. Red and yellow boxes are utilized to frame the ground truth position of the target initially given in the infrared and visible light images.

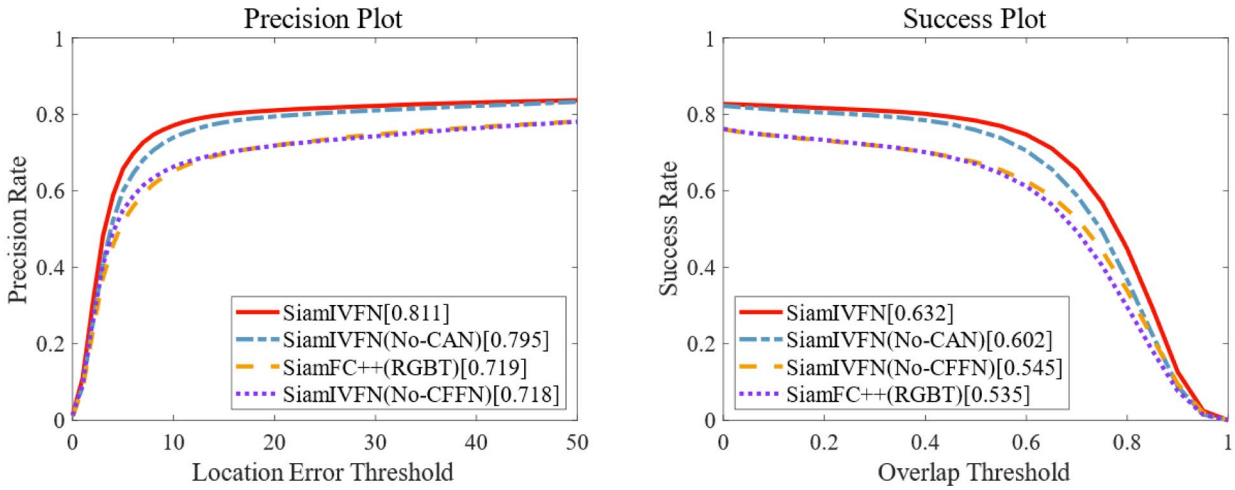
The second and third column image pairs are selected from *nightthreepeople* and *woman6*, whose background is complex. The complex background can easily interfere with the classification score of the object, making it impossible to distinguish the foreground and background correctly. The infrared image background is simple and easy to distinguish. The CFFN extracts the individual features of infrared and visible images, improving the stability of the tracker through the infrared port in complex backgrounds.

The false object was occluding the real object in *manwithbasketball twoperson*, the first and fourth columns. When the fake object pass by the real object and misalignment occurs, the infrared part of the real object is located in the visible part of the fake object, leading to the classification score of the fake object being higher than the real object, causing the tracker to make an error. Since the CFFN is a feature-level fusion, it can cope with slight misalignment.

To illustrate the effectiveness of the CAN, We separately selected 200 frames from the day and night sequences of the RGBT dataset. We visualize the contribution vectors for



(a) Comparison of visible light, infrared, and fusion tracking



(b) Prune experiments

Fig. 6 Comparison of visible light, infrared, and fusion tracking (**a**). Ablation experiment (**b**) on RGBT234

Table 2 Comparison of different coupling rates on RGBT234

Experiment (#)	Conv2	Conv3	Conv4	Precision rate	Success rate
1	0	0	0	71.8	54.5
2	0.25	0.50	0.75	**81.1**	**63.2**
3	0.25	0.75	0.50	72.9	52.3
4	0.50	0.25	0.75	72.5	54.3
5	0.50	0.75	0.25	66.5	47.6
6	0.75	0.25	0.50	70.9	50.1
7	0.75	0.50	0.25	64.0	45.1
8	1.00	1.00	1.00	71.9	53.5

The best precision and recall rates are listed in bold.

some frames in Fig. 8. The first 256 contribution vectors in the figure are calculated from visible light features, and the 257th–512th contribution vectors are calculated from infrared features. It can be found that in the nighttime sequences *beginhand*, the infrared feature has larger contribution weights (warm color). In contrast, in the daytime sequence *car*, the contribution weights of the visible light feature are relatively large. It means that the CAN pays more attention to features beneficial to the tracking task.

5.5 Efficiency analysis

We compare the efficiency of SiamIVFN with that of other fusion tracking methods in Fig. 9. It can be found that the speed of the proposed SiamIVFN greatly exceeds other fusion methods. SiamIVFN reached 147.6FPS, 124.6FPS

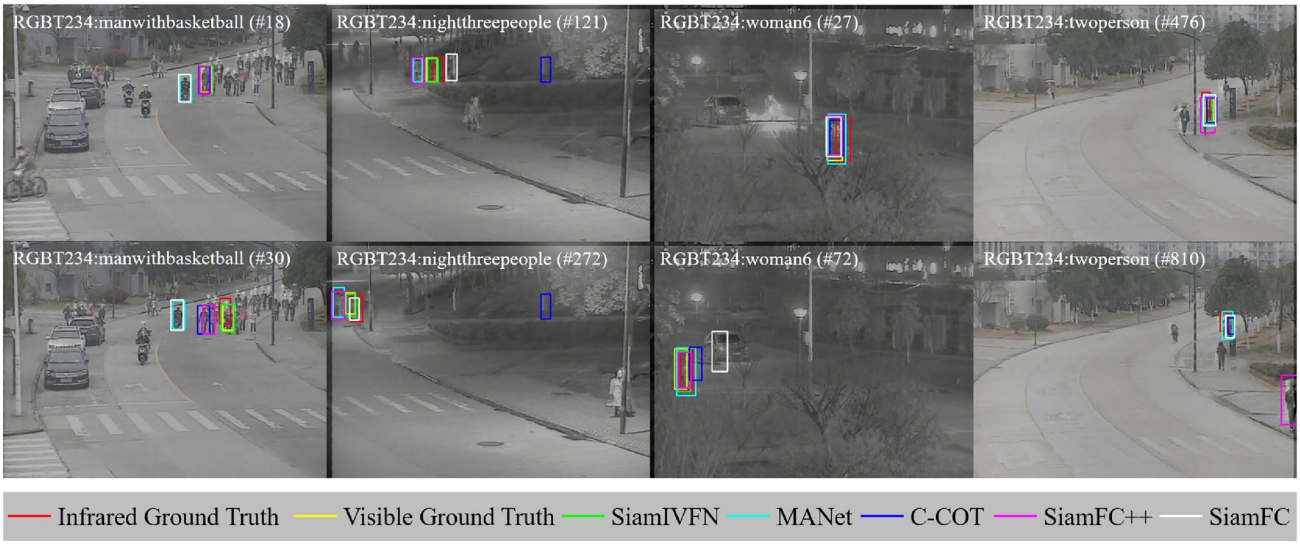


Fig. 7 Qualitative analysis of SiamIVFN and other trackers

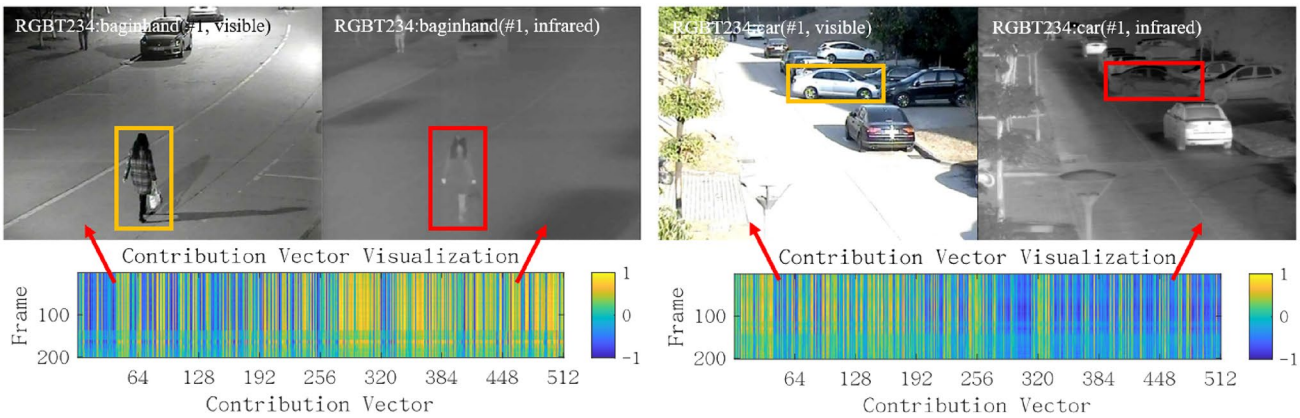
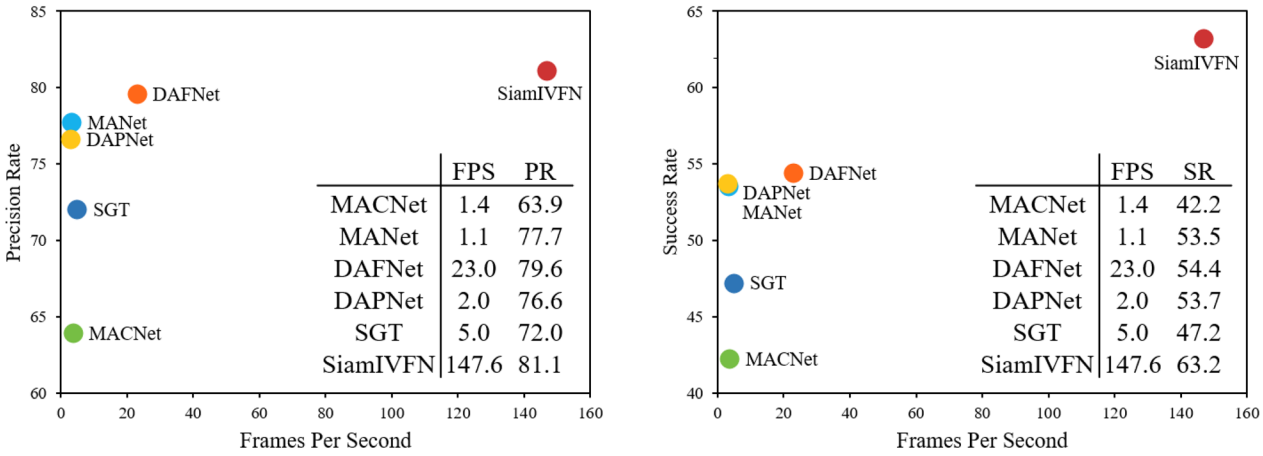


Fig. 8 Visualization of contribution vectors of CAN. The horizontal axis represents 512 vectors, and the vertical axis represents the number of frames in the video sequence. Color from cold to warm represents the value from – 1 to 1

faster than the second-best fusion tracker DAFNet. In the design of SiamIVFN, we give priority to speed and take the Siamese-based structure as the tracking head. Besides, both CFFN and CAN are more concise than the backbone of other fusion tracking methods.

Based on all the experiments performed in this section, we conclude that:

1. Compared with the visible tracking method (KCF, ECO, C-COT, MDNet, and SiameseFC) and the fusion tracking method (MANet, MACNet, SGT, DAPNet, and DAFNet), SiamIVFN achieves the best PR/SR score and the fastest tracking speed.
2. The performance of the fusion method is better than that of methods based on single-modal images, which shows the advantage of fusion.
3. With the two-stream structure and coupled filters used in CFFN, SiamIVFN can separately extract individual features and common features.
4. With CAN, which adaptively calculates the contributions of the infrared and visible light features, SiamIVFN is robust to various lighting conditions.



(a) Comparison of precision rate and speed

(b) Comparison of success rate and speed

Fig. 9 Speed comparison of various tracking methods

6 Conclusion

A novel RGB-T image-based tracking method, called SiamIVFN, is proposed in this paper. SiamIVFN can adaptively fuse the complementarity of infrared and visible light images to address the object tracking problem under various light conditions. SiamIVFN mainly contains two subnetworks, CFFN and CAN. Owing to the two-stream convolutional structure, CFFN can extract both common features and individual features from infrared and visible light image pairs. CFFN treats infrared and visible light images as complements of each other through the coupling filters. The common features of infrared and visible light images can be learned without additional computation. CFFN is a feature-level fusion network that can handle situations where visible light and infrared images are not rigorously aligned. Under various light conditions, CAN is designed to adaptively compute the contributions of different features, which could learn the weight coefficient of each channel through self-attention. Experiments performed on two RGB-T tracking benchmark datasets demonstrate that SiamIVFN outperforms other latest RGB-T tracking methods and can reach 147.6FPS. In the future, we will try to adopt other advanced architectures to let the network dynamically change the coupling rate, and combine temporal modalities to improve tracking performance.

Acknowledgements This research is sponsored by National Natural Science Foundation of China (62173143 and 61973122).

Data availability statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Liu T, Kong J, Jiang M, Liu C, Gu X, Wang X (2019) Collaborative model with adaptive selection scheme for visual tracking. Int J Mach Learn Cybern 10:215–228. https://doi.org/10.1007/s13042-017-0709-1
2. Ahmed I, Ahmad M, Ahmad A, Jeon G (2020) Top view multiple people tracking by detection using deep sort and yolov3 with transfer learning: within 5g infrastructure. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-020-01220-5
3. Zhou Z, Zhang W, Zhao J (2019) Robust visual tracking using discriminative sparse collaborative map. Int J Mach Learn Cybern 10:3201–3212. https://doi.org/10.1007/s13042-019-01011-7
4. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS (2016) Fully-convolutional Siamese networks for object tracking. In: Computer Vision—ECCV 2016 Workshops, pp 850–865. Springer, Cham.
5. Li B, Yan J, Wu W, Zhu Z, Hu X (2018) High performance visual tracking with Siamese region proposal network. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 8971–8980. https://doi.org/10.1109/CVPR.2018.00935
6. Li B, Wu W, Wang Q, Zhang F, Xing J, Yan J (2019) Siamrpn++: evolution of Siamese visual tracking with very deep networks. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 4277–4286. https://doi.org/10.1109/CVPR.2019.00441
7. Xu Y, Wang Z, Li Z, Yuan Y, Yu G (2020) Siamfc++: towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the AAAI Conference on Artificial Intelligence 34:12549–12556. https://doi.org/10.1609/aaai.v34i07.6944
8. Havens, K.J., Sharp, E.J. (2016) Chapter 3 - Remote sensing. In: Thermal Imaging Techniques to Survey and Monitor Animals in the Wild, pp 35–62. Academic Press, Boston. https://doi.org/10.1016/B978-0-12-803384-5.00003-8
9. Li C, Zhu C, Zheng S, Luo B, Tang J (2018) Two-stage modality-graphs regularized manifold ranking for rgb-t tracking. Signal Process Image Commun 68:207–217. https://doi.org/10.1016/j.image.2018.08.004

10. Li C, Wang X, Zhang L, Tang J, Wu H, Lin L (2017) Weighted low-rank decomposition for robust grayscale-thermal foreground detection. IEEE Trans Circ Syst Video Technol 27(4):725–738. https://doi.org/10.1109/TCSVT.2016.2556586
11. Wu A, Zheng W-S, Yu H-X, Gong S, Lai J (2017) Rgb-infrared cross-modality person re-identification. In: 2017 IEEE International Conference on computer vision (ICCV), pp 5390–5399. https://doi.org/10.1109/ICCV.2017.575
12. Xu D, Ouyang W, Ricci E, Wang X, Sebe N (2017) Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 4236–4244. https://doi.org/10.1109/CVPR.2017.451
13. Yun S, Choi J, Yoo Y, Yun K, Choi JY (2017) Action-decision networks for visual tracking with deep reinforcement learning. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 1349–1358. https://doi.org/10.1109/CVPR.2017.148
14. Yun X, Jing Z, Jin B (2016) Visible and infrared tracking based on multi-view multi-kernel fusion model. Opt Rev 23:244–253. https://doi.org/10.1007/s10043-015-0175-5
15. Zhang X, Ye P, Liu J, Gong K, Xiao G (2019) Decision-level visible and infrared fusion tracking via Siamese networks. In: The 9th Chinese Conference on information fusion, pp 742–750
16. Zhang X, Ye P, Qiao D, Zhao J, Peng S, Xiao G (2019) Object fusion tracking based on visible and infrared images using fully convolutional Siamese networks. In: 2019 22th International Conference on information fusion (FUSION), pp 1–8
17. Zhang X, Ye P, Leung H, Gong K, Xiao G (2020) Object fusion tracking based on visible and infrared images: a comprehensive review. Inform Fusion 63:166–187. https://doi.org/10.1016/j.inffus.2020.05.002
18. Liu J, Zhang S, Wang S, Metaxas D (2016) Multispectral deep neural networks for pedestrian detection. In: Wilson RC, Smith ERH, WAP (eds) Proceedings of the British Machine Vision Conference (BMVC), pp 73–17313. BMVA Press, York. https://doi.org/10.5244/C.30.73
19. Zhang P, Zhao J, Bo C, Wang D, Lu H, Yang X (2021) Jointly modeling motion and appearance cues for robust rgb-t tracking. IEEE Trans Image Process 30:3335–3347. https://doi.org/10.1109/TIP.2021.3060862
20. Li CL, Lu A, Zheng AH, Tu Z, Tang J (2019) Multi-adapter rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp 2262–2270. https://doi.org/10.1109/ICCVW.2019.00279
21. Zhang X, Ye P, Peng S, Liu J, Gong K, Xiao G (2019) Siamft: an rgb-infrared fusion tracking method via fully convolutional Siamese networks. IEEE Access 7:122122–122133. https://doi.org/10.1109/ACCESS.2019.2936914
22. Zhang X, Ye P, Peng S, Liu J, Xiao G (2020) Dsiammft: an rgb-t fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. Signal Process Image Commun 84:115756. https://doi.org/10.1016/j.image.2019.115756
23. Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W (2018) Distractor-aware Siamese networks for visual object tracking. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision—ECCV 2018. Springer, Cham, pp 103–119
24. Zhang Z, Peng H (2019) Deeper and wider Siamese networks for real-time visual tracking. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 4586–4595. https://doi.org/10.1109/CVPR.2019.00472
25. Li C, Liu L, Lu A, Ji Q, Tang J (2020) Challenge-aware rgbt tracking. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds) Computer Vision—ECCV 2020. Springer, Cham, pp 222–237
26. Yang R, Zhu Y, Wang X, Li C, Tang J (2019) Learning target-oriented dual attention for robust rgb-t tracking. In: 2019 IEEE International Conference on image processing (ICIP), pp 3975–3979. https://doi.org/10.1109/ICIP.2019.8803528
27. Wang C, Xu C, Cui Z, Zhou L, Zhang T, Zhang X, Yang J (2020) Cross-modal pattern-propagation for rgb-t tracking. In: 2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 7062–7071. https://doi.org/10.1109/CVPR42600.2020.00709
28. Wang Z, Chang S, Yang Y, Liu D, Huang TS (2016) Studying very low resolution recognition using deep networks. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 4792–4800. https://doi.org/10.1109/CVPR.2016.518
29. Li Y, Zhao H, Hu Z, Wang Q, Chen Y (2020) Ivfusenet: fusion of infrared and visible light images for depth prediction. Inform Fusion 58:1–12. https://doi.org/10.1016/j.inffus.2019.12.014
30. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
31. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
32. Huang L, Zhao X, Huang K (2021) Got-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Trans Pattern Anal Mach Intell 43(5):1562–1577. https://doi.org/10.1109/TPAMI.2019.2957464
33. Fan H, Lin L, Yang F, Chu P, Deng G, Yu S, Bai H, Xu Y, Liao C, Ling H (2019) Lasot: a high-quality benchmark for large-scale single object tracking. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 5369–5378. https://doi.org/10.1109/CVPR.2019.00552
34. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327. https://doi.org/10.1109/TPAMI.2018.2858826
35. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision—ECCV 2016. Springer, Cham, pp 499–515
36. Li C, Cheng H, Hu S, Liu X, Tang J, Lin L (2016) Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Trans Image Process 25(12):5743–5756. https://doi.org/10.1109/TIP.2016.2614135
37. Li C, Liang X, Lu Y, Zhao N, Tang J (2019) Rgb-t object tracking: benchmark and baseline. Pattern Recogn 96:106977. https://doi.org/10.1016/j.patcog.2019.106977
38. Henriques JF, Caseiro R, Martins P, Batista J (2015) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596. https://doi.org/10.1109/TPAMI.2014.2345390
39. Danelljan M, Bhat G, Khan FS, Felsberg M (2017) Eco: efficient convolution operators for tracking. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 6931–6939. https://doi.org/10.1109/CVPR.2017.733
40. Danelljan M, Robinson A, Shahbaz Khan F, Felsberg M (2016) Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Computer Vision—ECCV 2016, pp 472–488. Springer, Cham.
41. Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 4293–4302. https://doi.org/10.1109/CVPR.2016.465
42. Li C, Zhao N, Lu Y, Zhu C, Tang J (2017) Weighted sparse representation regularized graph learning for rgb-t object tracking. In: Proceedings of the 25th ACM International Conference on

multimedia, pp 1856–1864. Association for Computing Machinery, New York. https://doi.org/10.1145/3123266.3123289
43. Zhang H, Zhang L, Zhuo L, Zhang J (2020) Object tracking in RGB-T videos using modal-aware attention network and competitive learning. Sensors. https://doi.org/10.3390/s20020393
44. Zhu Y, Li C, Luo B, Tang J, Wang X (2019) Dense feature aggregation and pruning for RGBT tracking. In: Proceedings of the 27th ACM International Conference on multimedia, pp 465–472. Association for Computing Machinery, New York. https://doi.org/10.1145/3343031.3350928
45. Gao Y, Li C, Zhu Y, Tang J, He T, Wang F (2019) Deep adaptive fusion network for high performance rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp 91–99. https://doi.org/10.1109/ICCVW.2019.00017