IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2019, accepted August 17, 2019, date of publication August 22, 2019, date of current version September 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936914

SiamFT: An RGB-Infrared Fusion Tracking Method via Fully Convolutional Siamese Networks

XINGCHEN ZHANG[1], PING YE[1], SHENGYUN PENG[1,2], JUN LIU[1,3], KE GONG[1], AND GANG XIAO[1]

[1]School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China
[2]College of Civil Engineering, Tongji University, Shanghai 200092, China
[3]School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

Corresponding author: Gang Xiao (xiaogang@sjtu.edu.cn)

This work was sponsored in part by the National Program on Key Basic Research Project under Grant 2014CB744903, in part by the National Natural Science Foundation of China under Grant 61973212 and Grant 61673270, in part by the Shanghai Science and Technology Committee Research Project under Grant 17DZ1204304, in part by the Shanghai Industrial Strengthening Project under Grant GYQJ-2017-5-08.

ABSTRACT Object tracking based on visible images may fail when the visible images are unreliable, for example when the illumination condition is poor. Infrared images reveal thermal radiation of objects and are insensitive to these factors. Due to the complementary features of visible and infrared images, RGB-infrared fusion tracking has attracted widespread attention recently. In this paper, an RGB-infrared fusion tracking method based on the fully convolutional Siamese Networks, termed as SiamFT, is proposed. Visible and infrared images are firstly processed by two Siamese Networks, namely visible network and infrared network, respectively. Then, convolutional features of visible and infrared template images extracted from two Siamese Networks are concatenated to form fused template image. Convolutional features of visible and infrared search image are fused through the proposed feature fusion network adaptively. In particular, a modality weight computation method based on the response value of Siamese network is proposed to predict the reliability of different images. Cross-relation is then applied to the fused template feature and the fused search feature to produce the final response map, based on which the tracking results can be obtained. Extensive experiments indicate that the proposed SiamFT shows better performance than the-state-of-art fusion tracking algorithms at real-time speed.

INDEX TERMS Deep learning, fusion tracking, object tracking, Siamese network.

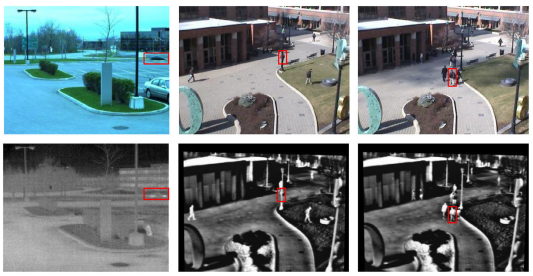
I. INTRODUCTION

Object tracking has received increasing attention in recent years due to its wide applications in many areas, such as robotics, surveillance, and human-machine interface. A lot of algorithms have been proposed to perform object tracking, among which the most popular ones are based on deep learning and correlation filter (CF). Tracking methods based on deep learning, especially the convolutional neural networks (CNN), can produce good tracking performance due to the strong feature representation ability of CNN. However, since the training and online update of the CNN model are time consuming, the CNN model is normally trained off-line and is kept fixed during tracking. In contrast, CF-based trackers can update model online and have a high frame rate due to the effective computation of correlation filters with fast Fourier transformation (FFT).

Currently, most video tracking algorithms are developed for visible images (RGB images) [1]. Despite remarkable progress, tracking algorithms based on visible images may fail as they may be unreliable in certain circumstance. For example, when the illumination conditions are poor. In contrast, infrared images reveal thermal information of objects and are insensitive to these factors. They can provide complementary information to visible images and show camouflaged objects under darkness etc., as shown in Fig. 1(a). Besides, in some situations, RGB images are more reliable than infrared images since they have color feature and can provide more details, as illustrated in Fig. 1(b). By leveraging the complementary information in visible and infrared images,

The associate editor coordinating the review of this article and approving it for publication was Yong Yang.

(a) Target in infrared images is more clear and distinguishable



(b) Target in visible images is more clear and distinguishable

FIGURE 1. Examples of complementary information in visible and infrared images [4].

the robustness of tracking algorithms can be significantly enhanced. As a consequence, in recent years, object tracking based on visible and thermal infrared images has become a hot research topic and is termed as RGB-infrared fusion tracking [2], [3].

Before deep learning and correlation filters, researchers performed fusion tracking with traditional techniques [3], [5], such as mean-shift and Camshift algorithms. These methods were not able to obtain good tracking performance when challenging factors present. Recently, both deep learning [6] and correlations filters [7] have been applied to fusion tracking. However, the tracking performance still needs to be improved. On one hand, some of the trackers can run at high speed, but the tracking precision is not good enough. For instance, Zhai *et al.* [8] proposed a fusion tracking method based on correlation filters whose frame rate was 224 frames per second (FPS). However, its tracking precision and robustness were not as good as the state-of-the-art trackers. On the other hand, some of these algorithms are very slow, although their tracking precisions are good. For example, the speed of the tracker proposed by Lan *et al.* [9] was 0.7 FPS, which was far from the real-time requirement. As a consequence, it is difficult to apply these trackers in practical scenarios.

In this paper, we propose an effective and also efficient RGB-infrared fusion tracking method, which can produce competitive tracking performance against the state-of-the-art trackers at real-time speed. Specifically, a fusion tracking method based on fully Siamese networks is proposed and is termed as SiamFT.

In summary, the main contributions of this paper are as follows:

- A fusion tracking method based on fully convolutional Siamese networks is proposed. To the best of our knowledge, this is the first work that performs RGB-infrared fusion tracking based on the Siamese networks by combining multimodal features. In particular, two Siamese networks are employed to process visible and infrared images respectively, which can exploit useful information more effectively.
- A modality weight computation method based on the response map of Siamese networks is proposed for the first time. The complementary features in visible and infrared images are then fused using these weights for better usage of multimodal information.
- Extensive experiments have been conducted on a variety of visible and infrared video pairs to verify the effectiveness of SiamFT in terms of both tracking precision and speed under challenging conditions.

The rest of the paper is organized as follows. Section II introduces some related work and Section III discusses the proposed fusion tracking algorithm. Then, experimental details and results are presented in Section IV and Section V, respectively. In Section VI some discussions are given and finally Section VII concludes the paper.

II. RELATED WORK

A. VISUAL OBJECT TRACKING

Video tracking has attracted a great amount of attention in recent years. Currently, two main kinds of methods in visual object tracking are based on deep learning [10] and correlation filter [11]. Methods based on deep learning mainly utilize its strong feature representation ability compared to handcrafted features. However, since the online update of deep learning model is time consuming, the deep tracking model is normally trained offline and is kept fixed during tracking. In contrast, CF-based methods are computationally efficient thus it can be updated online. However, the performance of CF-based trackers is slightly worse than deep learning-based ones [12].

A benchmark dataset is crucial in visual tracking. Wu *et al.* [13], [14] proposed visual tracking benchmarks (OTB) that greatly promote the development of object tracking. In addition, the VOT challenge [15] also provides a platform for the community to compare tracking performance under the same standard.

B. IMAGE FUSION

Image fusion aims to combine information from multiple images into a single image, which is able to provide a better data source for applications. Numerous image fusion algorithms have been proposed, which can be generally divided into pixel-level, feature-level and decision-level fusion approaches. Also, image fusion can either be performed in the spatial domain or transform domain. Before deep learning is introduced to image fusion community, the main image fusion methods include weighted average method [16], wavelet transform-based method [17], principal component analysis (PCA)-based method [18],

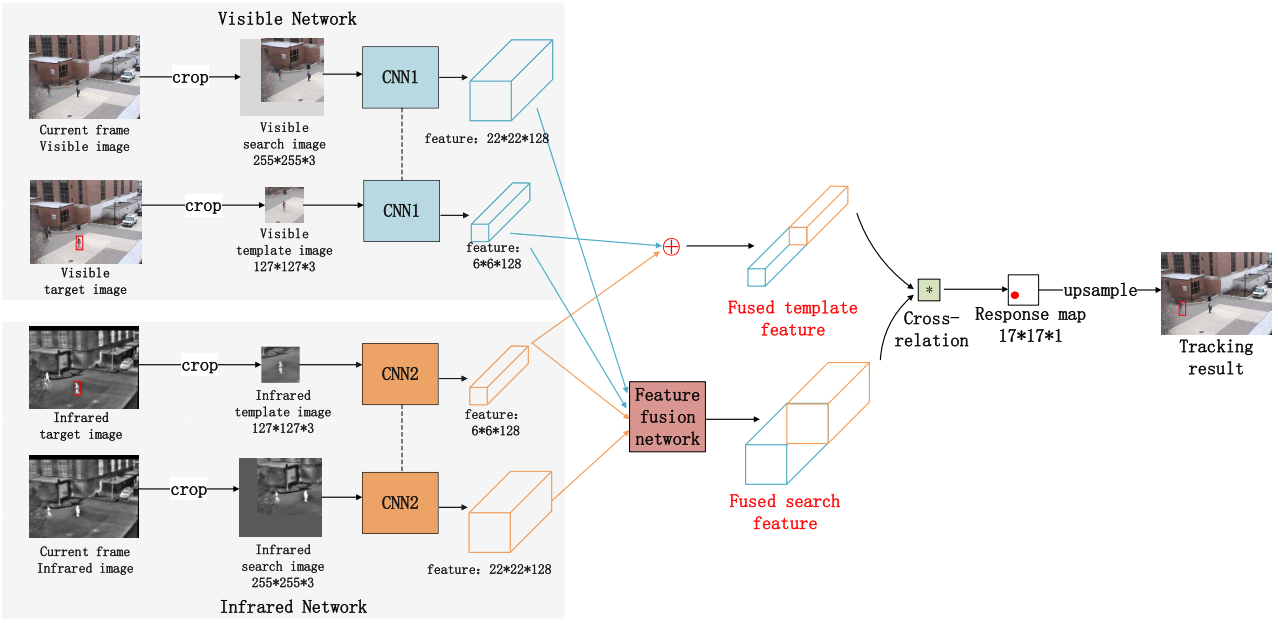


FIGURE 2. Flowchart of the proposed RGB-infrared fusion tracking algorithm based on Siamese networks.

sparse representation-based method [19] and compressive sensing-based method [20].

Recently, researchers began to perform image fusion based on deep learning methods [21], including multi-focus image fusion [22], medical image fusion [23], visible and infrared image fusion [24], [25], multi-exposure [26] image fusion etc. Regarding methods, CNN [27], generative adversarial networks (GAN) [28], Siamese networks [29], autoencoder [25] have been explored to conduct image fusion.

It should be mentioned that the aim of image fusion is different from that of fusion tracking. In image fusion, one normally aims to produce a fused image which has better visual quality. Whereas in fusion tracking only information about the target and its surroundings are important and thus should be extracted by tracking algorithms. The background information is not quite essential.

C. RGB-INFRARED FUSION TRACKING

In the last few years, fusion tracking has attracted a lot of interests and an increasing number of researches have been published in high-quality journals and well-known conferences [2], [6], [8], [9], [30]–[38]. Generally speaking, deep learning-based methods and CF-based methods are two main kinds of fusion tracking methods nowadays. Deep learning has shown its potentials in RGB tracking, thus researchers have started to apply deep learning to fusion tracking. For example, Xu *et al.* [30] presented a fusion tracking method based on CNN. A two-layer simple CNN was utilized there to perform fusion tracking, and the infrared channel was simply regarded as the fourth channel of the RGB image. Li *et al.* [6] proposed a two-stream CNN for fusion tracking, which utilized two CNNs to process visible and infrared images, respectively.

On the other hand, correlation filters have also produced promising performances in fusion tracking due their effectiveness and high efficiency. To the best of our knowledge, Wang *et al.* [7] presented the first CF-based fusion tracking work. After this, Zhai *et al.* [8] proposed an RGB-infrared tracking via cross-modal correlation filters. Yun *et al.* [38] also presented a CF-based fusion tracker. Although the research of fusion tracking based on correlation filters began in 2018, their highly competitive performances make them a promising research direction.

Apart from deep learning and CF-based methods, Li *et al.* [35] presented a graph-based fusion tracking approach. However, the frame rate of that method was only 8 FPS, which was far from the real-time requirement and thus was not suitable for practical usage.

III. PROPOSED METHODS

The novel aspects of the proposed method are discussed in this section. First, the network architecture is introduced. Then, the feature fusion network including a modality weight computation method is described.

A. NETWORK ARCHITECTURE

In the proposed method, features of visible and infrared images are firstly extracted and fused. Then, the fused features are utilized by a tracker to locate target. Compared to those methods which fuse images in pixel-by-pixel manner [2], [39], the proposed method mainly has two advantages. First, the proposed method fuses high-level features instead of pixels, thus it is more computationally efficient.

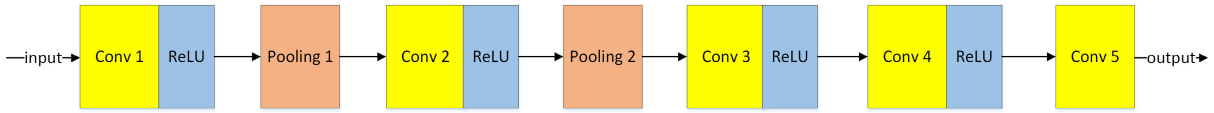


FIGURE 3. The network architecture of CNN in SiamFC [10] and SiamFT.

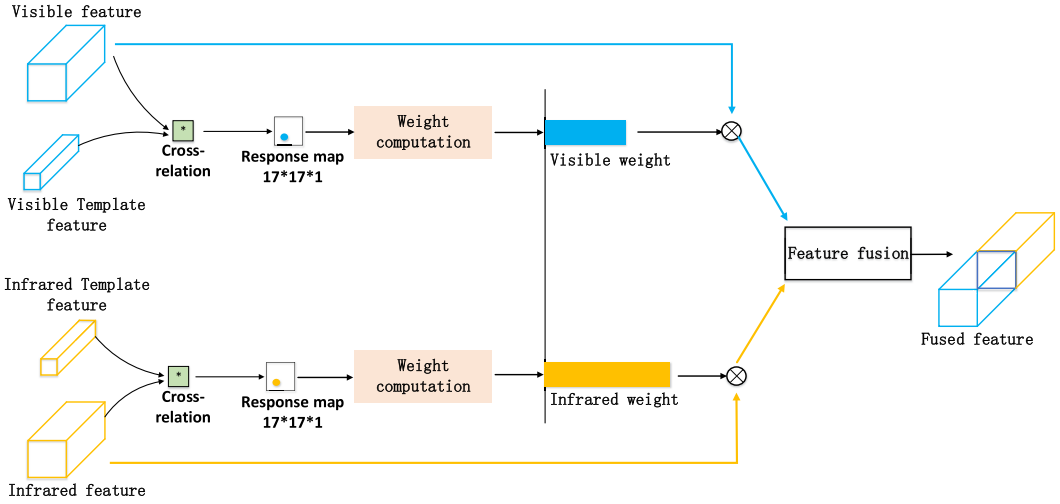


FIGURE 4. Flowchart of the proposed feature fusion network.

Second, the proposed method can produce effective feature representation directly which is crucial for tracking. Whereas the pixel-level fusion is easy to produce artifacts which may impair tracking performance.

The flowchart of our proposed approach, SiamFT, is illustrated in Fig. 2. Basically, two subnetworks are utilized in SiamFT, namely the visible network and the infrared network. They are used to process visible and infrared images, respectively. In this study, the SiamFC proposed by Bertinetto *et al.* [10] was utilized as the backbone due to its superior performance in both tracking precision and speed. Both the visible network and infrared network have the same architecture as the CNN part of SiamFC [10].

The structure of the CNN [10] is given in Fig. 3. As can be seen, a rectified linear unit (ReLU) layer is followed after each convolution layer except the last one. Also, pooling layers are merely utilized after the first two convolution layers. Besides, this CNN is fully convolutional hence there is no restrict requirement on the size of input images.

In SiamFT, the first frame containing target was chosen as the target image. In each network, the target image and current frame images were cropped into the template image and search images, respectively. Both the template and search images were centered at the tracking target object, and their sizes were $127 \times 127 \times 3$ and $255 \times 255 \times 3$, respectively. If the target was very close to the boundary, then we filled in the template and search image using mean pixel value after cropping. Then, the visible template and visible search images were fed to the two branches of the visible network to produce corresponding visible features. Similarly, the infrared template and infrared search images were fed to the two branches of the infrared network to produce corresponding infrared features. The visible template feature and the infrared template feature were then concatenated to produce a fused template feature. The visible search feature and the infrared search feature were fused through the proposed feature fusion network to produce a fused search feature. The proposed feature fusion network considered modality reliability, thus can adaptively and effectively fuse visible and infrared search features.

After obtaining the fused template feature and the fused search feature, the cross-relation between them was computed to obtain the response map which reflects the position of target. Finally, the position and size of the target in the current frame were computed by upsampling the response map.

It worth mentioning that the principle of the proposed framework is generic, thus other Siamese networks-based tracking methods can also be employed as the backbone. It should also be noted that due to the different imaging characters of visible and infrared images, the network which can effectively process them could be different. As a consequence, the CNN1 in the visible network and the CNN2 in the infrared network in Fig. 2 could be different.

Denote the CNN in visible network as φ, the CNN in infrared network as φ', visible search image as x_v, visible template image as z_v, infrared search image as x_t, infrared template image as z_t, then the response map of the proposed

Algorithm 1: Fusion Tracking Algorithm Based on Siamese Networks (SiamFT)

Input: Registered visible images and infrared images, groundtruth of the 1st frame
Output: Predicted position and size of object in each frame
***Initialization*:**
Crop the visible target image to obtain visible template image z_v
Crop the infrared target image to obtain infrared template image z_t
Tracking:
for *each frame i* **do**
    Crop current frame visible image to obtain x_v
    Crop current frame infrared image to obtain x_t
    Feed z_v and x_v into the visible network to obtain $\varphi(z_v)$ and $\varphi(x_v)$
    Feed z_t and x_t into the infrared network to obtain $\varphi'(z_t)$ and $\varphi'(x_t)$
    Compute visible modality weight based on z_v and x_v
    Computer infrared modality weight based on z_t and x_t
    Fuse $\varphi(z_v)$ and $\varphi'(z_t)$ using the feature fusion network, to obtain fused template feature $\varphi(z_v) \oplus \varphi'(z_t)$
    Fuse $\varphi(x_v)$ and $\varphi'(x_t)$ using the feature fusion network, to obtain fused search feature $\varphi(x_v) \oplus \varphi'(x_t)$
    Compute the response map according to equation (1)
    Upsample the response map to obtain the predicted position of target
end

fusion tracking method is:

$$responseMap = (\varphi(z_v) \oplus \varphi'(z_t)) * (\varphi(x_v) \oplus \varphi'(x_t)), \quad (1)$$

where $*$ indicates the correlation operation, $\oplus$ indicates feature fusion. Note that in this work, the cross-relation was implemented using convolution efficiently. Also, we aimed to give a proof-of-concept study thus we utilized the same network for CNN1 and CNN2. Therefore, in the present study, φ and φ' were identical. The algorithm of SiamFT is illustrated in Algorithm 1.

B. FEATURE FUSION NETWORK

Effective feature fusion is crucial in RGB-infrared fusion tracking. We designed a feature fusion network as shown in Fig. 4

1) MODALITY WEIGHT COMPUTATION

The key of the proposed feature fusion network is the computation of modality weights which reflect the reliability of different images and are crucial for efficiently leveraging complementary information of multimodal images. In this study, we proposed a modality weight computation method based on the response value obtained from the cross-relation computation.

From experiments, we observed that if an image was reliable, normally it had two characters. First, it had similar features with the template image. This can be reflected by the response value. To be more specific, in the proposed SiamFT, we computed two cross-relations in each frame, i.e. for both visible and infrared images, thus two response values were generated. We observed that if the visible image was more reliable (for example light condition was good or no background clutter), the visible response value was higher. Otherwise, if the infrared image was more reliable (for example when the illumination condition was poor), the infrared response value was higher. Second, the object did not move too fast in two consecutive frames normally. Therefore, the distance between the predicted target location in two consecutive frames can also reflect the reliability of image (or modality). Based on these observations, we proposed computing the modality weight as:

$$weight_i = \begin{cases} max(R_i), & if\ d < threshold \\ \dfrac{max(R_i)}{\sqrt{d}}, & if\ d \geq threshold, \end{cases} \quad (2)$$

where R is the response value computed from cross-relation and d is the distance between the predicted target location in two consecutive frames. The subscript i indicates modality, namely i is t for infrared images and is v for visible images. *threshold* can be chosen based on experiments.

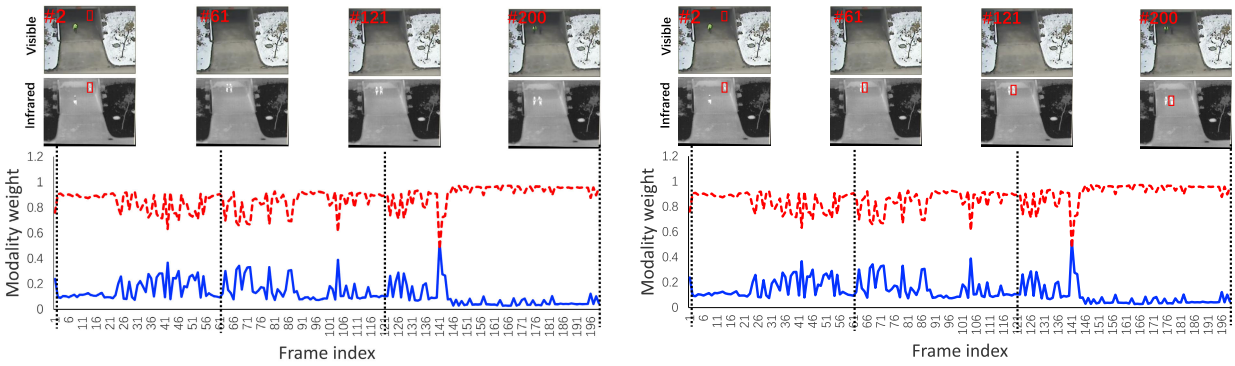
After computing weights of both visible and infrared images according to (2), we normalized the weights as:

$$\omega_v = \frac{weight_v}{weight_v + weight_t}, \quad (3)$$

$$\omega_t = \frac{weight_t}{weight_v + weight_t}, \quad (4)$$

where the subscript v and t indicate visible and infrared images, respectively.

Two examples of modality weights computed using the proposed method are presented in Fig. 5. In the first one, the target (car) is clear in the first frame and gradually becomes unclear due to over-exposure. As can be seen, in frame 3 the visible image is more clear than the infrared one, therefore the visible weight has a larger value. In frame 32, as the car turns gradually, it is becoming exposure. As a consequence, the reliability of visible image decreases and the infrared weight increases. In frame 100, the car light is too bright that it is difficult to see the car in the visible image. In contrast, one can see the car clearly in the infrared image since it is insensitive to illumination change. Therefore, in frame 100 the infrared image has a much higher weight. In the second example, it is difficult to see the target from visible images due to darkness, whereas it is easy to locate the target in infrared images. It can be seen clearly that the computed modality weights indeed indicate different reliability degrees of two modalities. Besides, the proposed modality weight computation method can adapt



(a) Target in visible image is not clear due to over-exposure

(b) Target in visible image is not clear due to darkness

FIGURE 5. Illustration of modality weights based on modality response value. The red dash line and blue solid line indicate the weights of infrared and visible modalities, respectively. In both over-exposure and darkness conditions, the proposed method can adaptively predict the modality reliability.

to the change of scene. These examples demonstrate the effectiveness of the proposed modality weight computation method in predicting the reliability of visible and infrared images.

2) FEATURE FUSION

In this study, the template was obtained from the first frame and was not updated during tracking, therefore the fused template feature was

$$\varphi(z_v) \oplus \varphi'(z_t) = concat(\varphi(z_v), \varphi'(z_t)), \tag{5}$$

where *concat* is concatenating two feature vectors into a single feature vector. Note that the modality weights were not applied in template features which were computed based on the first frame. Therefore, multiplication operation is not needed in equation (5).

The fused search feature was computed as

$$\varphi(x_v) \oplus \varphi'(x_t) = concat(\omega_v \times \varphi(x_v), \omega_t \times \varphi'(x_t)), \tag{6}$$

where ω_v and ω_t are the modality weight of visible and infrared images, respectively. Note that the modality weights were updated in each frame starting from the second frame.

The cross-relation of these two fused features was then performed to obtain the response map.

IV. EXPERIMENTS

To test the performance of the proposed SiamFT, a lot of experiments were conducted using various aligned visible and infrared video pairs. In this section, the implementation details, sequences, and compared trackers, as well as evaluation metrics are introduced.

A. IMPLEMENTATION DETAILS

In this study, all experiments were conducted using a desktop equipped with an NVIDIA GTX 1080Ti GPU and i7-8700K CPU.

As mentioned previously, we utilized the same CNN for both visible and infrared network. Therefore, we trained a Siamese network and applied it to both modalities. The Siamese network was trained using the ImageNet dataset [40], which contains a large number of annotated videos. Basically, we firstly generated training pairs from ImageNet dataset. A training pair consists of a template image and corresponding search image, as described in Section III-A. The template image and search image were extracted from two frames of a video. They were at most T frames apart and both contained the object.

For each pair of the template and search images fed to the network, a response map D was obtained. The loss function was defined based on the labeled value $y[u]$ and ground-truth value $v[u]$ of each individual element $u \in D$. Specifically, the loss function was constructed as

$$L(y, v) = \frac{1}{|D|} \sum_{u \in D} log(1 + exp(-y[u] \cdot v[u])). \tag{7}$$

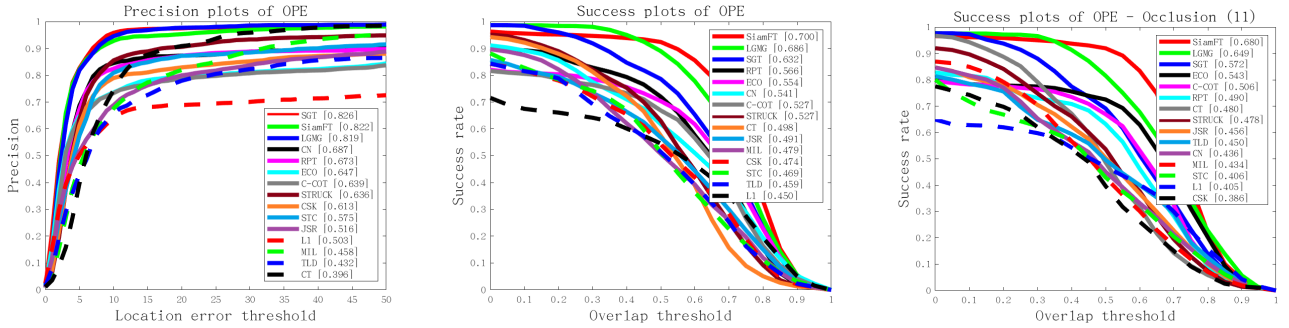
The parameters of the Siamese network θ were obtained by applying Stochastic Gradient Descent (SGT) to minimize the loss function in (7). The training was performed for 50 epochs, each consisting of 50,000 sampled pairs. The learning rate was annealed geometrically at each epoch from 10^{-2} to 10^{-5}.

More details on the training can be found in [10].

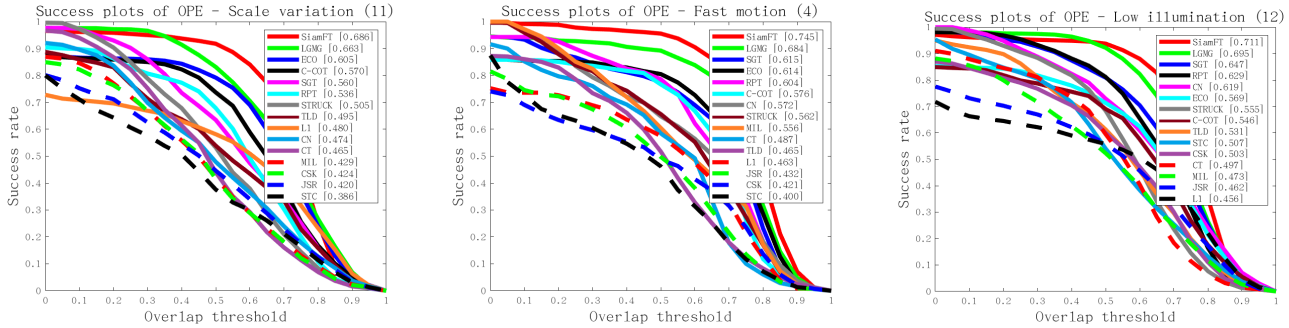
B. SEQUENCES AND COMPARED TRACKERS

Nineteen RGB-infrared video pairs were utilized for testing performance. These videos were well aligned and covered a variety of challenging conditions, such as heavy occlusion and fast motion. More details about the attributes are listed in Table 1.

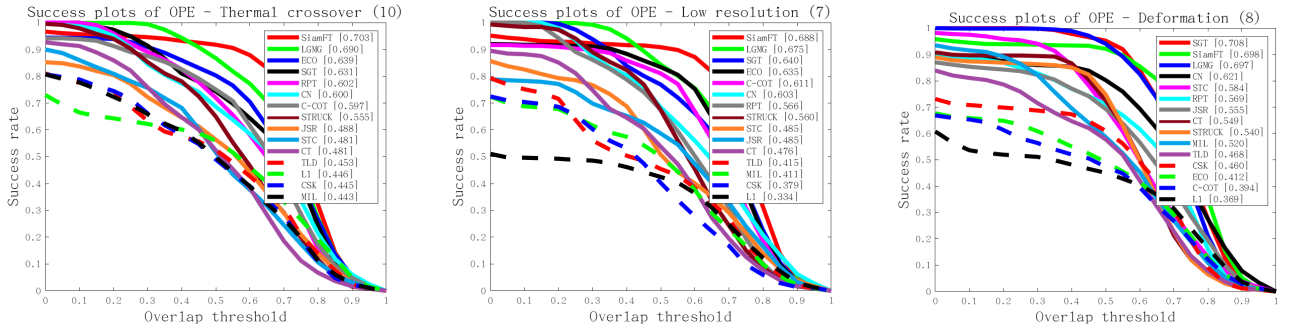
The performance of SiamFT on nineteen sequences are compared with 14 state-of-the-art trackers, including ECO [41], C-COT [42], CN [43], JSR [3], CSK [44], CT [45], L1 [46], MIL [47], RPT [48], STC [49], STRUCK [50], TLD [51], SGT [33], LGMG [36]. In these methods, the JSR, L1, SGT and LGMG methods are designed for RGB-infrared fusion tracking, while the others are originally developed for tracking based on visible images. To investigate the effect



(a) Overall precision plot (b) Overall success plot (c) Success plot of occlusion



(d) Success plot of scale variation (e) Success plot of fast motion (f) Success plot of low illumination



(g) Success plot of thermal crossover (h) Success plot of low resolution (i) Success plot of deformation

FIGURE 6. Comparison of tracking results in terms of precision rate and success rate. The number in the title denotes the number of videos with this attribute.

TABLE 1. Attribute information included in experiments.

Attribute	Description	Video numbers
OCC	Occlusion	11
SV	Scale Variation	11
FM	Fast Motion	4
LI	Low Illumination	12
TC	Thermal Crossover	10
LR	Low Resolution	7
DEF	Deformation	8

of multimodal information fusion, ECO and C-COT only run with visible images. Regarding other methods, the multimodal version of these trackers can be obtained according to [33]. Some results of these trackers on these video sequences can be obtained from [33]. The threshold in (2) was set to 5.

C. EVALUATION METRICS

In this study, we utilized the success rate (SR) and precision rate (PR) to evaluate fusion tracking performance [13]. Success means that the overlapping between the predicted bounding box and groundtruth box is larger than a threshold, where the overlapping is defined as:

$$O(a, b) = \frac{|a \bigcap b|}{|a \bigcup b|}, \tag{8}$$

where a and b indicates the predicted bounding box and groundtruth, respectively. The success plot shows the trends of success rate when the threshold changes from 0 to 1. The area under curve (AUC) is employed to rank different methods effectively.

Precision means that the center location error (CLE) between the predicted bounding box and the groundtruth is

TABLE 2. Success rate (SR %) on 19 sequences. The best three results are shown in red, green and blue, respectively. Best viewed in color.

	ECO	C-COT	CN	CSK	CT	MIL	RPT	STC	STRUCK	TLD	SGT	LGMG	JSR	L1	Ours
BlueCar	75.8	71.6	39.8	39.6	33.8	39.3	63.6	27.2	36.8	48.0	49.9	72.5	38.9	62.1	73.2
BusScale	79.2	76.7	50.2	50.4	45.7	48.2	57.0	44.5	46.3	52.7	60.8	73.7	53.1	71.2	76.4
BusScale1	71.5	62.6	41.8	43.0	42.5	39.1	65.3	40.1	40.2	68.3	44.5	66.0	46.7	64.9	69.6
Crossing	79.5	75.0	77.1	77.8	65.7	79.1	77.8	60.8	78.0	31.0	78.7	77.6	62.4	64.1	83.5
CrowdNig	65.3	67.4	78.6	20.9	10.5	20.9	44.1	43.1	47.8	13.5	76.5	70.8	65.4	74.5	68.7
Exposure4	74.4	73.5	54.1	19.3	35.3	28.0	63.5	34.4	55.2	23.3	64.0	76.7	56.1	49.2	75.3
FastMotor	60.8	55.4	39.8	47.7	41.1	47.0	34.7	26.8	40.9	1.0	42.9	52.3	40.1	2.0	42.1
GarageHover	35.5	34.3	63.1	72.9	63.3	43.2	71.7	55.9	57.0	38.1	67.1	69.2	49.9	10.8	76.9
Jogging	39.7	32.5	74.7	69.2	56.8	60.3	75.1	63.7	59.5	63.0	74.7	76.7	62.1	25.3	74.7
LightOcc	63.7	59.8	49.1	40.3	26.9	42.3	48.8	25.1	44.7	43.7	44.7	52.9	42.1	45.2	68.9
Minibus1	5.5	5.5	5.5	53.7	50.2	53.0	5.9	5.2	51.5	46.0	58.0	66.7	51.7	66.5	67.1
MinibusNig	66.5	68.5	57.6	55.6	52.7	54.1	66.6	53.9	52.9	61.8	61.5	72.1	31.9	72.6	69.7
Otcbvs1	76.0	77.8	66.1	61.0	63.1	71.4	69.9	67.3	61.6	53.2	75.0	76.1	75.8	13.8	83.0
Quarreling	26.7	24.0	62.6	10.7	68.7	61.8	51.4	46.8	65.5	63.1	72.7	70.7	29.6	13.6	72.5
RainyCar1	72.3	68.0	53.5	59.4	53.3	7.0	67.1	48.3	56.8	67.1	56.3	62.5	5.2	7.8	69.4
Torabi	4.4	4.6	4.2	4.0	54.8	39.1	3.8	58.0	3.9	9.8	62.7	57.1	59.9	5.7	63.5
Tricycle	72.9	69.8	62.1	68.4	59.9	69.4	69.2	65.6	65.9	59.7	73.2	73.3	67.7	68.1	76.0
tunnel	13.0	13.1	80.0	71.2	60.3	57.4	73.6	59.2	72.2	66.9	77.3	74.1	31.5	78.2	48.8
WalkingNig1	69.1	61.7	67.6	61.9	61.5	64.0	65.6	76.5	64.4	69.3	60.3	63.0	70.2	74.1	70.6
Overall	55.4	52.7	54.1	47.4	49.8	47.9	56.6	46.9	52.7	45.9	63.2	68.6	49.1	45.0	70.0

TABLE 3. Precision rate (PR %) on 19 sequences. The best three results are shown in red, green and blue, respectively. Threshold is 5 pixels. Best viewed in color.

	ECO	C-COT	CN	CSK	CT	MIL	RPT	STC	STRUCK	TLD	SGT	LGMG	JSR	L1	Ours
BlueCar	67.8	63.5	56.5	49.8	20.0	41.7	54.8	64.4	44.4	32.2	82.6	68.7	60.9	79.1	96.5
BusScale	52.0	49.0	45.5	45.0	13.5	31.0	49.0	43.0	22.5	16.5	48.0	49.0	36.0	69.5	56.5
BusScale1	60.0	54.1	54.1	80.0	57.7	15.3	67.1	83.5	54.1	67.1	74.1	70.6	63.5	57.7	70.6
Crossing	87.7	77.7	65.4	57.8	2.3	70.0	73.1	86.2	60.0	1.5	56.9	48.7	21.5	14.6	96.2
CrowdNig	95.8	97.6	100.0	28.1	11.4	28.1	67.1	79.6	88.6	21.0	100	100	93.4	100.0	89.2
Exposure4	99.3	96.6	71.4	24.5	21.1	28.6	97.3	53.1	76.9	25.2	88.4	98.0	86.4	72.1	97.3
FastMotor	100.0	100.0	37.0	97.0	89.0	74.0	58.0	19.0	44.0	1.0	100	100	18.0	2.0	62.0
GarageHover	29.5	40.2	77.3	88.8	47.4	10.8	95.6	49.0	61.8	17.1	96.0	94.4	43.0	10.8	98.8
Jogging	60.6	60.1	95.0	87.0	43.6	62.0	90.4	73.7	63.0	37.2	99.7	99.7	61.2	72.3	98.7
LightOcc	68.4	65.2	71.6	61.2	2.8	16.0	66.4	50.4	61.6	66.4	63.6	66.4	59.2	60.0	84.0
Minibus1	5.4	5.4	4.1	44.6	44.6	44.6	40.5	40.5	31.1	37.8	91.9	93.2	36.5	64.9	66.2
MinibusNig	58.1	63.8	93.3	67.6	59.1	47.6	69.5	81.0	71.4	95.2	67.6	70.5	36.2	60.0	83.8
Otcbvs1	100	100	92.6	77.3	27.0	96.9	81.0	90.8	58.9	22.7	100	100	100	19.6	100
Quarreling	34.7	32.4	74.4	13.2	70.0	50.7	44.3	50.2	81.7	79.5	93.2	90.4	27.4	22.4	89.5
RainyCar1	100	100	100	100	63.3	6.7	90.0	5.0	100	83.3	100	98.3	5.0	5.0	98.3
Torabi	12.5	12.5	12.5	12.5	6.7	2.1	0.4	30.4	1.3	2.9	10.0	11.3	26.7	2.5	17.5
Tricycle	95.4	95.4	65.4	68.5	77.7	95.4	83.9	69.2	94.6	71.5	99.2	99.2	86.2	43.9	100
tunnel	14.0	13.5	99.5	74.5	11.5	50.0	87.5	60.5	93.5	42.5	98.5	98.0	20.0	99.0	61.5
WalkingNig1	100	98.7	100	98.7	84.0	98.0	100	100	98.7	99.3	100	99.3	99.3	100	95.3
Overall	64.7	63.9	68.7	61.3	39.6	45.8	67.3	57.5	63.6	43.2	82.6	81.9	51.6	50.3	82.2

smaller than a chosen threshold. The precision plot shows the trends when the threshold changes from small value to large value. The threshold is set to 5 pixels in this work as the targets are relatively small in the above videos. Note that in this work, the visible image is the primary modality, therefore the groundtruth of visible images is chosen to compute success rate and precision score.

V. RESULTS

This section compares the results of the proposed method against the state-of-the-art trackers by presenting both quantitative and qualitative results.

A. QUANTITATIVE RESULTS

The quantitative results of our proposed method and the compared trackers on each sequence in terms of SR and PR are shown in Table 2 and 3, respectively. In summary, the proposed SiamFT achieves the best results in terms of SR and the second best results (slightly worse than the best one) in terms of PR among all compared trackers on the nineteen RGB-infrared videos. In particular, SiamFT outperforms all compared trackers in 7 sequences in terms of SR and in 6 sequences in terms of PR. Besides, SiamFT stays in the rank of top 3 in 16 videos in terms of SR and in 13 videos in terms of PR. This clearly demonstrates the effectiveness of SiamFT in RGB-infrared fusion tracking.

The effectiveness of SiamFT is also indicated by the precision and success plots presented in Fig. 6. As can be seen, the proposed SiamFT not only achieves very competitive overall performances in terms of both PR and SR, it outperforms almost all compared trackers in terms of SR in most examined challenging scenarios, such as occlusion, scale variation, and fast motion. The only exception is that SiamFT performs slightly worse than SGT when the target has large deformation. The experimental results clearly indicate that the proposed SiamFT is less sensitive to different challenging conditions and thus can provide robust performance. This is because that infrared images can provide thermal features

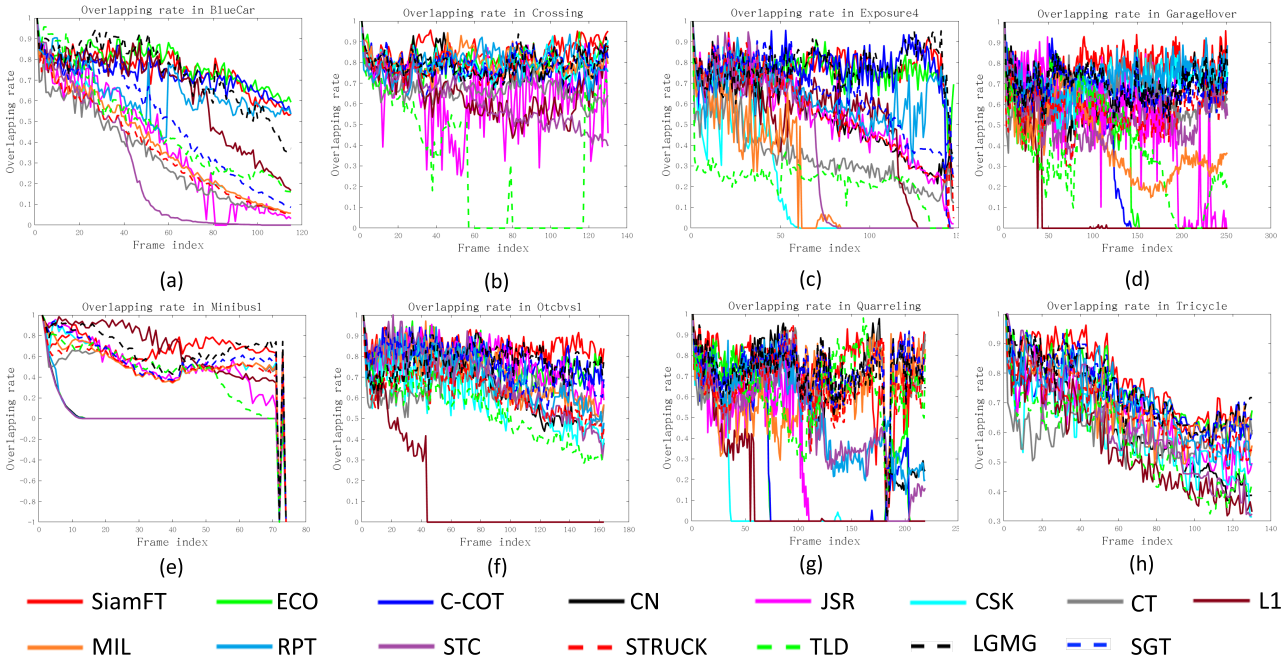


FIGURE 7. Frame-by-frame comparison of 15 trackers on 8 challenging sequences in terms of overlapping rate. The vertical axis and the horizontal axis indicate overlapping rate and frame index, respectively. These 8 sequences are *BlueCar, Crossing, Exposure4, GarageHover, Minibus1, Otcbvs1, Quarreling, Tricycle*.

which are more robust in some challenging situations, while visible images can provide better features in some other scenarios. The proposed method can effectively utilize complementary information in visible and infrared images to form better feature representation. In addition, the modality weights enable a more reliable modality to contribute more to the fused feature.

Besides, we can observe from Table 2 and 3 that, SGT, LGMG, and the proposed SiamFT outperform ECO and C-COT with a very clear margin, although ECO and C-COT have both achieved very competitive performances in tracking based on visible images. This indicates that integrating visible and infrared data can boost tracking performance. The improvements are more significant while encountering certain challenges, i.e., low illuminations and thermal crossover, demonstrating the complementary benefits from visible and infrared images.

B. QUALITATIVE RESULTS

Fig. 7 gives the qualitative frame-by-frame comparison of trackers in terms of overlapping rate in some videos. As can be seen, the proposed tracker achieves a relatively higher overlapping rate and runs stably in these videos, demonstrating the stability of the proposed SiamFT. Fig. 8 shows the qualitative comparison of bounding boxes among different trackers on some videos, which visually demonstrates the effectiveness of our approach. It can be found that the proposed tracker is more robust under some challenging conditions, such as occlusion (e.g. *Exposure4, Minibus1*), scale variation (e.g. *Quarreling*), poor illumination condition (e.g. *LightOcc*).

C. RUNNING TIME COMPARISON

The running speed of the proposed tracker is around 28-32 FPS thus can meet real-time requirement. This is much faster and more practically useful than methods whose speed are only several FPS, such as the SGT (5 FPS) and LGMG (7 FPS), demonstrating the efficiency of the proposed method.

VI. DISCUSSION

This section gives some discussions on the proposed method and results.

A. CHALLENGE-BASED PERFORMANCE

The original aim of utilizing infrared images in tracking is to improve performance when visible images are not reliable. For example, when the illumination condition is poor. Surprisingly, experimental results show that the proposed method outperforms ECO and C-COT, which are among the best trackers using visible images, under all examined challenging conditions with a clear margin. We believe that this is because infrared images can provide complementary information to visible images. The proposed SiamFT can effectively leverage complementary features to improve tracking performance under adverse challenging scenarios.

B. INFRARED-SPECIFIC NETWORK

In this study, the networks were trained with ImageNet, which consists of visible images only. Although the proposed SiamFT already achieves the best overall results among compared trackers with real-time speed, we think that by fine-tuning the infrared network in SiamFT with infrared images

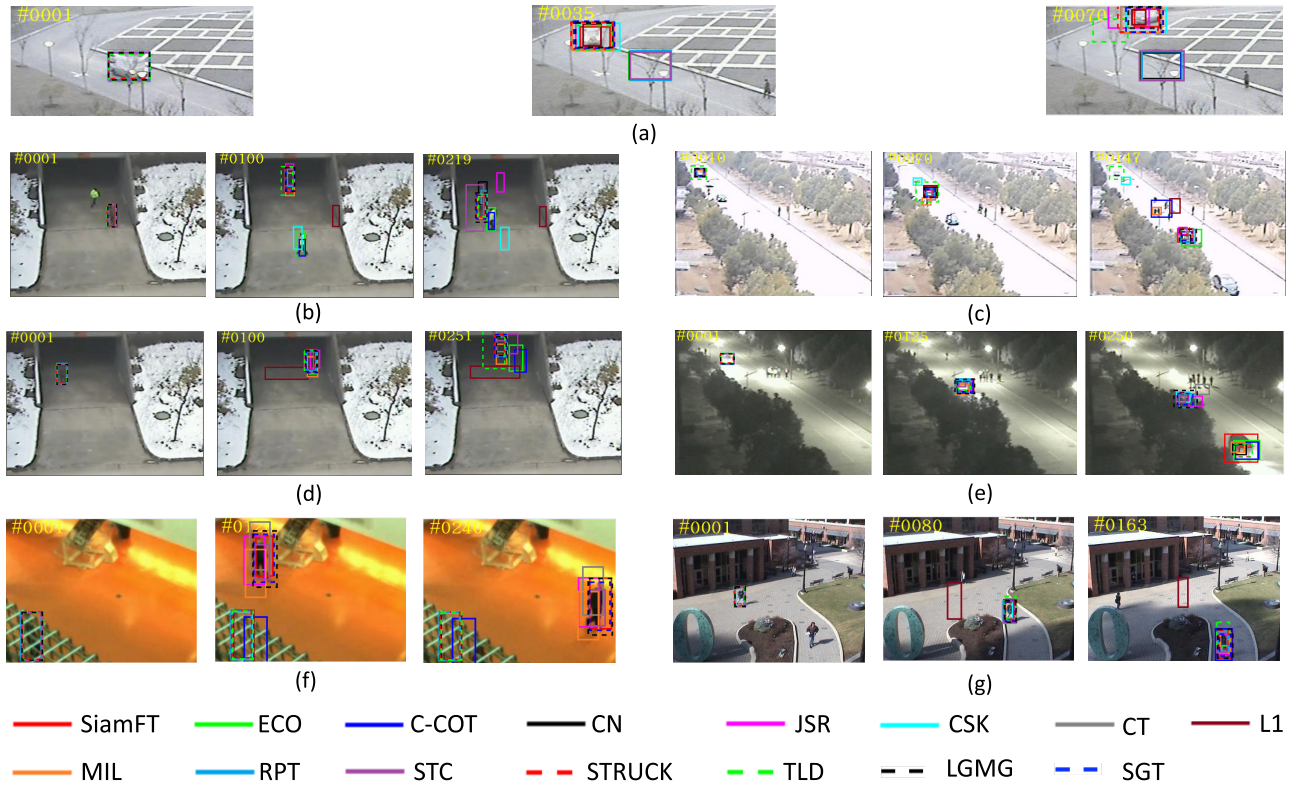


FIGURE 8. Qualitative comparison of 15 trackers on some videos under some challenging situations, such as occlusion (e.g. ***Exposure4, Minibus1***), scale variation (e.g. ***Quarreling***), poor illumination condition (e.g. ***LightOcc***). (a) ***Minibus1***, (b) ***Quarreling***, (c) ***Exposure4***, (d) ***GarageHover***, (e) ***LightOcc***, (f) ***Torabi*** (g) ***Otcbvs1***.

may improve its performance further. This is because that visible and infrared images have different imaging characteristics, thus might require different networks to process them more effectively.

C. FEATURE FUSION NETWORK

Fully exploiting complementary information in visible and infrared images is crucial for the success of fusion tracking. To this end, we designed the feature fusion network to fuse features of different modalities effectively by considering the reliability of different images during tracking. Although the proposed method can predict the reliability of different modalities, it depends on the manual selection of threshold value. In future, we will explore a more intelligent method to achieve this. For example, we can try to design a subnetwork to directly learn the modality weight.

VII. CONCLUSION

In this paper, a fusion tracking method using visible and infrared images based on the fully convolutional Siamese Networks, termed as SiamFT, is proposed. To the best of our knowledge, this is the first time that Siamese network is utilized to perform fusion tracking by combining multimodal features. Specifically, two Siamese networks, namely visible network and infrared network, are employed to process visible and infrared images respectively. A feature fusion network is designed to adaptively fuse the visible and infrared features extracted from two Siamese networks. In addition, a modality weight computation method is proposed to predict the reliability of visible and infrared images. Extensive experiments on challenging sequences demonstrate that the proposed approach achieves better performance than the state-of-the-art trackers. Besides, the proposed approach is effective and can run at real-time speed.

REFERENCES

[1] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.

[2] X. Zhang, G. Xiao, P. Ye, D. Qiao, J. Zhao, and S. Peng, "Object fusion tracking based on visible and infrared images using fully convolutional siamese networks," in *Proc. 22nd Int. Conf. Inf. Fusion*, to be published.

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

[4] J. W. Davis and V. Sharma, "Fusion-based background-subtraction using contour saliency," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Sep. 2005, p. 11.

[5] G. Xiao, X. Yun, and J. Wu, "A new tracking approach for visible and infrared sequences based on tracking-before-fusion," *Int. J. Dyn. Control*, vol. 4, no. 1, pp. 40–51, Mar. 2016.

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

[7] Y. Wang, C. Li, and J. Tang, "Learning soft-consistent correlation filters for RGB-T object tracking," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Cham, Switzerland: Springer, 2018, pp. 295–306.

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

[9] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust RGB-infrared tracking system," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9887–9897, Dec. 2019.
[10] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 850–865.
[11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[12] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1328–1338.
[13] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[14] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[15] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.
[16] H. Yin, "Tensor sparse representation for 3-D medical image fusion using weighted average rule," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2622–2633, Nov. 2018.
[17] P. Hill, M. E. Al-Mualla, and D. Bull, "Perceptual image fusion using wavelets," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1076–1088, Mar. 2017.
[18] C. He, Q. Liu, H. Li, and H. Wang, "Multimodal medical image fusion based on IHS and PCA," *Proc. Eng.*, vol. 7, pp. 280–285, Jan. 2010.
[19] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.
[20] T. Wan and Z. Qin, "An application of compressive sensing for image fusion," *Int. J. Comput. Math.*, vol. 88, no. 18, pp. 3915–3930, Dec. 2011.
[21] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.
[22] X. Yan, S. Z. Gilani, H. Qin, and A. Mian, "Unsupervised deep multi-focus image fusion," 2018, *arXiv:1806.07272*. [Online]. Available: https://arxiv.org/abs/1806.07272
[23] K.-J. Xia, H.-S. Yin, and J.-Q. Wang, "A novel improved deep convolutional neural network model for medical image fusion," *Cluster Comput.*, vol. 22, pp. 1515–1527, Jan. 2019.
[24] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," 2018, *arXiv:1804.08992*. [Online]. Available: https://arxiv.org/abs/1804.08992
[25] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2018.
[26] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732.
[27] H. Hermessi, O. Mourali, and E. Zagrouba, "Convolutional neural network-based multimodal image fusion via similarity learning in the shearlet domain," *Neural Comput. Appl.*, vol. 30, no. 7, pp. 2029–2045, Oct. 2018.
[28] J. Y. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2018.
[29] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, p. 1850018, 2018.
[30] N. Xu, G. Xiao, X. Zhang, and D. P. Bavirisetti, "Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences," in *Proc. 4th Int. Conf. Virtual Reality*, Feb. 2018, pp. 44–49.
[31] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for RGB-infrared tracking," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 7008–7015.
[32] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for RGB-infrared object tracking," *Pattern Recognit. Lett.*, to be published.
[33] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.
[34] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1856–1864.
[35] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. ECCV*, Sep. 2018, pp. 808–823.
[36] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, "Learning local-global multi-graph descriptors for RGB-T object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[37] X. Zhang, P. Ye, J. Liu, K. Gonge, and G. Xiao, "Decision-level visible and infrared fusion tracking via siamese networks," in *Proc. 9th Chin. Conf. Inf. Fusion*, to be published.
[38] X. Yun, Y. Sun, X. Yang, and N. Lu, "Discriminative fusion correlation learning for visible and infrared tracking," *Math. Problems Eng.*, vol. 2019, May 2019, Art. no. 2437521.
[39] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, and C. N. Canagarajah, "The effect of pixel-level fusion on object tracking in multi-sensor surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
[41] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6638–6646.
[42] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 472–488.
[43] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.
[44] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 702–715.
[45] D. Wang and H. Lu, "Visual tracking via probability continuous outlier model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3478–3485.
[46] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proc. 14th Int. Conf. Inf. Fusion*, Jul. 2011, pp. 1–8.
[47] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.
[48] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 353–361.
[49] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
[50] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.
[51] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 864–877.

XINGCHEN ZHANG received the B.Sc. degree from the Huazhong University of Science and Technology, in 2012, and the Ph.D. degree from the Queen Mary University of London, in 2017. He is currently a Postdoctoral Research Fellow with the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, China. He is also the Director of the Artificial Intelligence and Image Processing Group, Advanced Avionics and Intelligent Information (AAII) Laboratory. His current research interests include object fusion tracking, image fusion, deep learning, and computer vision.

PING YE is currently pursuing the Ph.D. degree with the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, China. She is also a member of the Advanced Avionics and Intelligent Information (AAII) Laboratory. Her major research interests include multi-source heterogeneous image fusion and object tracking based on deep learning.

SHENGYUN PENG is currently pursuing the bachelor's degree in civil engineering with Tongji University, China. He is also a Research Student with the Advanced Avionics and Intelligent Information (AAII) Laboratory, Shanghai Jiao Tong University. His current research interests include computer vision, object tracking, video object segmentation, FPGA development, and portable hardware devices.

JUN LIU is currently pursuing the master's degree with the School of Automation and Information Engineering, Sichuan University of Science and Engineering, China. He is also a Research Student with the Advanced Avionics and Intelligent Information (AAII) Laboratory, School of Aeronautics and Astronautics, Shanghai Jiao Tong University. His current research interests include visual object tracking and information fusion.

KE GONG is currently pursuing the master's degree with the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, China. He is also a member of the Advanced Avionics and Intelligent Information (AAII) Laboratory. His current research interests include object detection, deep learning, and power line detection.

GANG XIAO received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Full Professor with the School of Aeronautics and Astronautics and the Director of the Advanced Avionics and Intelligent Information (AAII) Laboratory, Shanghai Jiao Tong University. His current research interests include image fusion, target tracking, and avionics integration and simulation.