Signal Processing: Image Communication 84 (2020) 115756

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion[☆]

Xingchen Zhang[a], Ping Ye[a], Shengyun Peng[b], Jun Liu[c], Gang Xiao[a,*]

[a] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, 200240, China
[b] College of Civil Engineering, Tongji University, Shanghai, 200062, China
[c] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin, 644000, China

ARTICLE INFO

Keywords:
Object tracking
RGB-T fusion tracking
Dynamic Siamese networks
Deep learning
Image fusion

ABSTRACT

The task of object tracking is very important since its various applications. However, most object tracking methods are based on visible images, which may fail when visible images are unreliable, for example when the illumination conditions are poor. To address this issue, in this paper a fusion tracking method which combines information from RGB and thermal infrared images (RGB-T) is presented based on the fact that infrared images reveal thermal radiation of objects thus providing complementary features. Particularly, a fusion tracking method based on dynamic Siamese networks with multi-layer fusion, termed as DSiamMFT, is proposed. Visible and infrared images are firstly processed by two dynamic Siamese Networks, namely visible and infrared network, respectively. Then, multi-layer feature fusion is performed to adaptively integrate multi-level deep features between visible and infrared networks. Response maps produced from different fused layer features are then combined through an elementwise fusion approach to produce the final response map, based on which the target can be located. Extensive experiments on large datasets with various challenging scenarios have been conducted. The results demonstrate that the proposed method shows very competitive performance against the-state-of-art RGB-T trackers. The proposed approach also improves tracking performance significantly compared to methods based on images of single modality.

1. Introduction

Object tracking has received great attention in recent years due to its applications in various areas, such as entrance system, robotics, and transportation management. Various trackers have been designed, among which the most popular ones are based on deep learning [1] and correlation filters (CF) [2]. Most of these tracking algorithms are developed for tracking based on visible images [3]. However, these trackers may fail when visible images are not reliable, for instance when the illuminations are poor. Contrarily, infrared images are insensitive to these factors because they reveal thermal information of objects.

On one hand, infrared images can provide complementary benefits with visible images and show objects when they are not clear in visible images as shown in Fig. 1(a). On the other hand, RGB images are more reliable than infrared images in some situations since they can provide more details like color feature and texture information, as illustrated in Fig. 1(b). Therefore, it would be helpful to fuse complementary information from visible and infrared images in tracking.

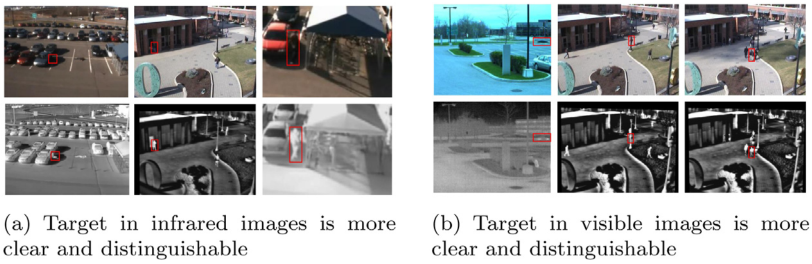
In recent years, object tracking based on visible and infrared images has become an active research topic, and is termed as RGB-Thermal (RGB-T) tracking [6] or fusion tracking [5,7]. Both deep learning methods [8] and correlations filters [9] have been applied to fusion tracking, which significantly improve the performance of fusion tracking compared to traditional methods [7,10]. However, the tracking performance still needs to be improved. The major problem of current methods is that they fail to strike a good balance between precision and speed. Some trackers can run at a high speed, whereas the tracking precisions are not good enough. Some trackers have good precisions but they are very slow. For instance, Zhai et al. [11] proposed a CF-based fusion tracking method whose frame rate was 224 FPS (frame per second). However, its tracking precision was not good enough. Lan et al. [12] proposed a fusion tracking algorithm with high precision but the speed was only 0.7 FPS, which was far from the real time requirement. It is difficult to apply these slow trackers in practical applications.

In this study, we aim to bridge the above-mentioned gap by proposing a fast fusion tracking method which can produce competitive tracking performance against the state-of-the-arts. In particular, an

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.image.2019.115756.

[*] Corresponding author.
E-mail address: xiaogang@sjtu.edu.cn (G. Xiao).

https://doi.org/10.1016/j.image.2019.115756
Received 15 April 2019; Received in revised form 9 November 2019; Accepted 17 December 2019
Available online 5 February 2020


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

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

Fig. 1. Examples of complementary information in visible and infrared images [4,5].

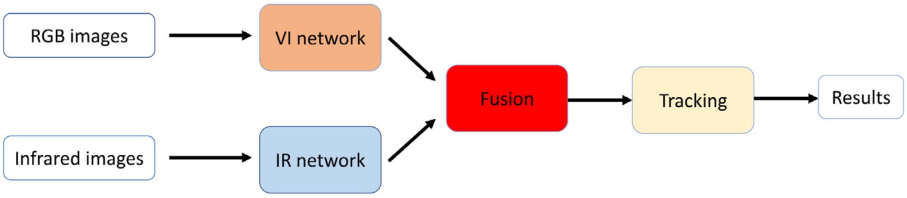


Fig. 2. Feature-level fusion tracking framework.

RGB-T tracker based on dynamic Siamese networks [13] is proposed and is termed as DSiamMFT.

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

- An RGB-T fusion tracking method based on dynamic Siamese networks is proposed. To the best of our knowledge, this is the first work that performs fusion tracking based on dynamic Siamese networks.
- Two dynamic Siamese networks are employed to process visible and infrared images respectively, which can exploit multi-modal information more effectively. The complementary features of visible and infrared images from multiple layers of the network are fused to produce better results.
- Extensive experiments have been conducted on large-scale visible and infrared image datasets to verify the significance of the proposed method in terms of tracking precision and speed.

The rest of the paper is organized as follows. Section 2 introduces some related work and Section 3 discusses the proposed fusion tracking algorithm. Then, experimental details and results are presented in Sections 4 and 5, respectively. In Section 6 some discussions are given and finally Section 7 concludes the paper.

2. Related work

2.1. Image fusion

Image fusion aims to fuse information from multiple images into a more informative single image. A lot of algorithms have been proposed, which can be generally divided into pixel-level, feature-level and decision-level fusion methods. Image fusion methods mainly contain weighted average method, wavelet-based, principal components analysis (PCA)-based, sparse representation-based and deep learning-based methods.

After deep learning is introduced into the field of image fusion recently, researchers have performed different image fusion tasks based on deep learning [14], such as multi-focus image fusion [15], medical image fusion [16], visible and infrared image fusion [17], multi-exposure image fusion [18]. Various deep learning methods, including CNN [19], Generative Adversarial Networks (GAN) [20], Siamese networks [21], autoencoder [17] have been explored to conduct image fusion.

2.2. RGB-T fusion tracking

Researches on fusion tracking based on visible and infrared images began more than ten years ago. However, the research was limited since the lack of large-scale datasets. The research of RGB-T tracking was boosted until comprehensive RGB-T fusion tracking datasets are available recently, such as RGBT210 [22]. In the past three years, an increasing number of RGB-T tracking algorithms have been published in high quality journals or well-known conferences [8,11,12,22–27].

Generally speaking, RGB-T fusion tracking algorithms can be divided into five categories according to their adopted theories, namely traditional methods, sparse representation (SR)-based, graph-based, correlation filter-based and deep learning-based approaches. Among these methods, deep learning-based ones are most popular in recent years. This is mainly because that deep learning model can learn effective features which are crucial in object tracking. For example, Xu et al. [28] presented a method based on CNN. In that work, a two-layer simple CNN was utilized to perform fusion tracking, and the infrared channel was regarded as the fourth channel of the RGB image. Li et al. [8] proposed a two-stream CNN for fusion tracking, which utilized two CNNs to process visible and infrared images, respectively.

3. Methods

3.1. RGB-T fusion tracking via siamese networks

A typical feature-level fusion tracking framework is illustrated in Fig. 2. Normally, a CNN is employed in the visible (VI) and another CNN is used in the infrared (IR) network [8]. However, current methods still have difficulties in making a balance between precision and speed.

Siamese networks have been widely applied to visual tracking since 2016 because of their simple structure, good performance and efficiency [1,29–33]. However, Siamese networks have been rarely applied in RGB-T fusion tracking so far. In this study, we propose employing dynamic Siamese networks to perform RGB-T fusion tracking, aiming at striking a good balance between precision and speed. In particular, two fully-convolutional Siamese networks are employed to process visible and infrared images, respectively. Therefore, the similarity function can be computed for all translated sub-windows within the search image in one evaluation, which leads to relatively high tracking speed. Besides, due to the strong feature representation ability of CNN, the learned background suppression transformation and target appearance variation transformation, as well as the multi-layer feature fusion, the proposed method obtains good tracking performance.

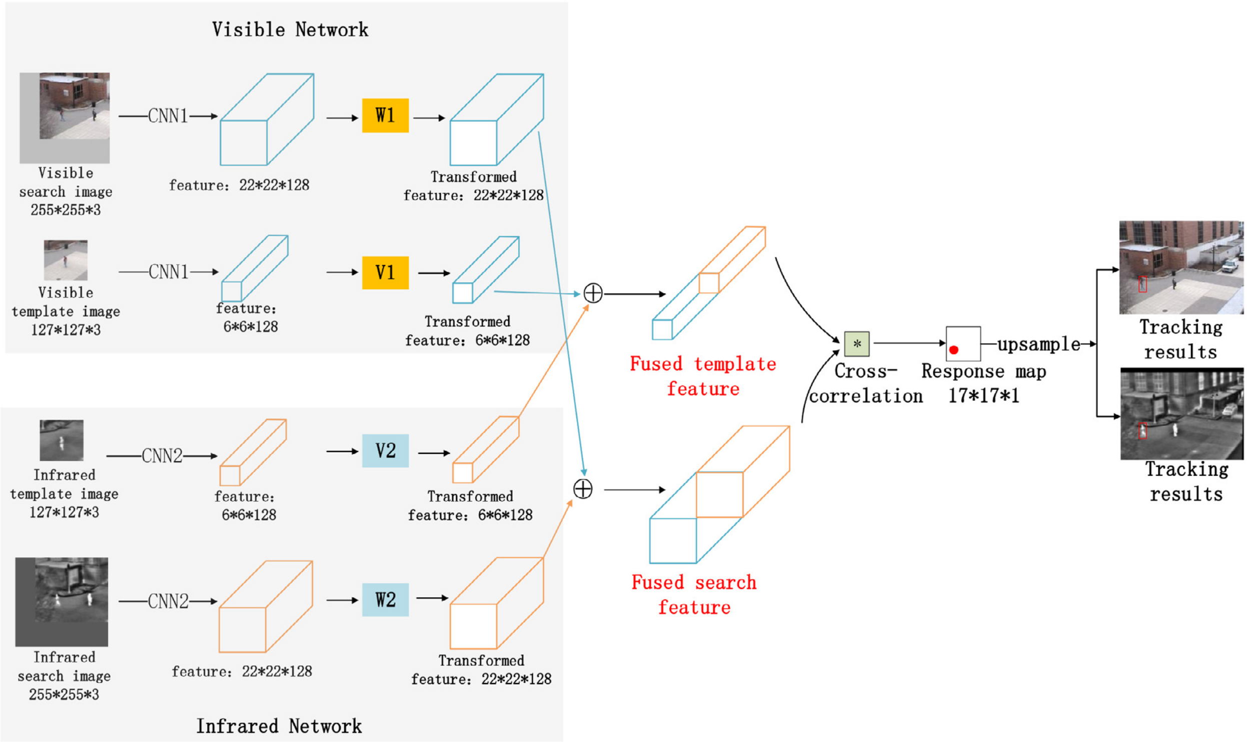


Fig. 3. Flowchart of the proposed fusion tracking algorithms based on dynamic Siamese networks.

3.2. Network architecture of the RGB-T tracker based on dynamic siamese networks

In this study, the DSiam proposed by Guo et al. [13] is utilized as the backbone of the proposed fusion tracking framework due to its good performance in both tracking precision and speed. However, it should be mentioned that the principle of the proposed framework is generic, thus other Siamese network-based tracking methods can also be employed.

The flowchart of fusion tracking based on dynamic Siamese networks (DSiamFT) is illustrated in Fig. 3. Basically, two dynamic Siamese networks are utilized in DSiamFT, namely the visible network and infrared network. Since each Siamese network has two branches, thus in our network there are four branches in total. In each network, the target image (in this study, the first frame containing target is chosen) and current frame images are cropped into template image and search images respectively. Both the template and search images are centered at the tracking target object, and their size are 127 × 127 × 3 and 255 × 255 × 3, respectively. If the target is very close to the boundary, then one needs to fill in the image using mean pixel value after cropping. Then, the template and search image are fed into two branches of a Siamese network to produce corresponding features.

Compared to SiamFC [1], the dynamic Siamese networks are equipped with two online-learned transformations, namely the target appearance variation transformation (denoted as **V**) and the background suppression transformation (denoted as **W**) [13]. Denote the CNN in visible network as φ, the CNN in infrared network as φ', current frame visible search image as x_v, the first frame visible search image as $x_{v,1}$, visible template image as $z_{v,1}$, current frame infrared search image as x_t, the first frame infrared search image as $x_{t,1}$, infrared template image as $z_{t,1}$, then the transformed features are:

$$f^5_{xvt} = \mathbf{V1} * \varphi(x_v), \tag{1}$$

$$f^5_{zvt} = \mathbf{W1} * \varphi(z_{v,1}), \tag{2}$$

$$f^5_{xtt} = \mathbf{V2} * \varphi'(x_t), \tag{3}$$

$$f^5_{ztt} = \mathbf{W2} * \varphi'(z_{t,1}), \tag{4}$$

where $f^5_{xvt}, f^5_{zvt}, f^5_{xtt}, f^5_{ztt}$ are transformed visible search feature, transformed visible template feature, transformed infrared search feature and transformed infrared template feature, respectively. $*$ denotes circular convolution that can be efficiently solved in frequency domain. The superscript 5 indicates the 5th layer. The CNN in each Siamese network has 5 layers, thus the 5th layer is the last layer. **V1** and **V2** denote the appearance variation transform in visible and infrared images, respectively. **W1** and **W2** are the background suppression transform, respectively.

Both transformations **V** and **W** are learned online [13]. To be more specific, regularized linear regression (RLR) is used to calculate **V** and **W** [13]. Here we discuss the transformation learning using the visible Siamese network as example. The same idea also applies to the infrared Siamese network.

The appearance variation transformation **V1** is used to update the deep features of target template $z_{v,1}$. It aims to encourage the $f^5(z_{v,1})$ being similar to $f^5(z_{v,i-1})$, thus it is learned from the 1st frame and $(i-1)$th frame by considering temporally smooth variation of the target. Specifically, after obtaining the tracking result at the $(i-1)$th frame, we have the target z_{i-1}. It is assumed that the target variation is temporally smooth. Then, we get the target appearance variation transformation **V1** by

$$\mathbf{V1} = \mathop{\arg\min}_{\mathbf{V1}} \|\mathbf{V1} * \varphi(z_{v,1}) - \varphi(z_{v,i-1})\|^2 + \lambda_v \|\mathbf{V1}\|^2, \tag{5}$$

where λ_v controls the regularization degree. Thanks to the desirable property of circular convolution $*$ [34], from Eq. (5) we can efficiently obtain **V1** by

$$\mathbf{V1} = \mathscr{F}^{-1}\left(\frac{\mathscr{F}^{\star}(\varphi(z_{v,1})) \odot \mathscr{F}(\varphi(z_{v,i-1}))}{\mathscr{F}^{\star}(\varphi(z_{v,1})) \odot \mathscr{F}(\varphi(z_{v,1})) + \lambda_v}\right), \tag{6}$$

where $\mathscr{F}$ is the discrete Fourier transformation (DFT), $\mathscr{F}^{-1}$ is the inverse DFT, $\star$ indicates complex-conjugate, $\odot$ denotes the elementwise multiplication.

The background variation transformation **W1** is utilized to update the deep features of search image $x_{v,1}$. It aims to highlight the deep feature of target neighborhood regions and alleviate the interference of irrelevant background features. It is learned based on the $(i-1)$th frame and its Gaussian version. After tracking at the $(i-1)$th frame, we have the target location and can crop the image to region $g_{v,i-1}$ centering at the target location and with the same size of the template image $z_{v,1}$. Then we multiply $g_{v,i-1}$ with a Gaussian weight map to get $\overline{g}_{v,i-1}$ to properly highlight the target regions. We can then get the background variation transformation **W1** by

$$\mathbf{W1} = \arg\min_{\mathbf{W1}} \|\mathbf{W1} * \varphi(g_{v,i-1}) - \varphi(\overline{g}_{v,i-1})\|^2 + \lambda_w \|\mathbf{W1}\|^2, \tag{7}$$

where λ_w is the parameter which controls the regularization degree. Similarly, we can efficiently have

$$\mathbf{W1} = \mathscr{F}^{-1}\left(\frac{\mathscr{F}^{\star}(\varphi(g_{v,i-1})) \odot \mathscr{F}(\varphi(\overline{g}_{v,i-1}))}{\mathscr{F}^{\star}(\varphi(g_{v,i-1})) \odot \mathscr{F}(\varphi(g_{v,i-1})) + \lambda_w}\right). \tag{8}$$

The target variation and background suppression transformations **V1** and **W1** enables the original Siamese network [1] with proper online adaptation ability. The corresponding transformations **V2** and **W2** for the infrared Siamese network can be learned similarly. More details about the computation of these transformations can be found in [13].

The transformed visible template feature and transformed infrared template feature are fused to produce the fused template feature. Similarly, the transformed visible search feature and transformed infrared search feature are fused to obtain fused search feature. The next step is to compute the cross-correlation between these two fused features. By doing so one can obtain the response map which reflects the position of target. The response map of the fusion tracking method proposed in this study is:

$$\mathbf{S}^5 = (f^5_{zvt} \oplus f^5_{ztt}) \otimes (f^5_{xvt} \oplus f^5_{zvt}), \tag{9}$$

where $\oplus$ indicates feature fusion, $\otimes$ indicates the correlation operation on two fused feature tensors. Also, feature fusion is achieved through concatenation as a proof of concept. Finally, the position of the target in current frame can be obtained by upsampling the response map. The algorithm of DSiamFT is illustrated in Algorithm 1.

Algorithm 1: DSiamFT

```
Input: Registered visible 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 $z_{v,1}$
Crop the infrared target image to obtain infrared template image $z_{t,1}$
Tracking:
for each frame i do
    Crop current frame images to obtain $x_v$ and $x_t$
    Compute V1 using $z_{v,1}$ and $z_{v,i-1}$
    Compute V2 using $z_{t,1}$ and $z_{t,i-1}$
    Compute W1 using $x_{v,i-1}$
    Compute W2 using $x_{t,i-1}$
    Feed $z_v$ and $x_v$ into the visible network to obtain $f^5_{zvt}$ and $f^5_{xvt}$
    Feed $z_t$ and $x_t$ into the infrared network to obtain $f^5_{ztt}$ and $f^5_{xtt}$
    Fuse $f^5_{xvt}$ and $f^5_{xtt}$ to obtain fused search feature $f^5_{xvt} \oplus f^5_{xtt}$
    Fuse $f^5_{zvt}$ and $f^5_{ztt}$ to obtain fused template feature $f^5_{zvt} \oplus f^5_{ztt}$
    Compute the response map $\mathbf{S}^5$ according to Eq. (9)
    Upsample the response map to obtain the predicted position of
      target
end
```

Note that due to the different characteristics of visible and infrared images, the network which can effectively process them and extract features should be different. As a consequence, the CNNs in visible network and infrared network should be different. However, in this work our aim is to demonstrate the principle of the proposed method, thus we use the same CNN in both VI and IR network. The CNN is pretrained using ImageNet [35] provided by Bertinetto et al. [1].

Table 1
Attribute information of RGBT210 dataset [22].

Attribute	Description	Attribute	Description
NO	No occlusion	DEF	Deformation
PO	Partial occlusion	FM	Fast motion
HO	Heavy occlusion	SV	Scale variation
LI	Low illumination	MB	Motion blur
LR	Low resolution	CM	Camera moving
TC	Thermal crossover	BC	Background clutter

3.3. RGB-T tracker based on dynamic siamese networks with multi-layer fusion

In DSiamFT, only features of the last layer are utilized. Actually, one could fuse multi-layer features to obtain better results. In this study, we propose a method to utilize multi-layer features in fusion tracking via dynamic Siamese networks, which is then termed as DSiamMFT. The flowchart of DSiamMFT is shown in Fig. 4. In DSiamMFT, features from 4th and 5th layers are utilized. Note that the target appearance variation (denoted as $\mathbf{V1}'$ and $\mathbf{V2}'$) and background suppression transformation (denoted as $\mathbf{W1}'$ and $\mathbf{W2}'$) are also applied to corresponding branches in both VI and IR network.

Then, the transformed feature from the 4th layer of visible network is fused with the transformed feature from the 4th layer of infrared network, to produce the fused template feature $f^4_{zvt} \oplus f^4_{ztt}$ and fused search feature $f^4_{xvt} \oplus f^4_{xtt}$. Here, the superscript 4 indicates the 4th layer. Cross-correlation is then computed for these two fused features to produce a response map $\mathbf{S}^4$. Response maps $\mathbf{S}^4$ and $\mathbf{S}^5$ are then fused using an elementwise weight map $\boldsymbol{\Omega}^l$ (l indicates the layer), yields

$$\mathbf{S} = \boldsymbol{\Omega}^4 \odot \mathbf{S}^4 + \boldsymbol{\Omega}^5 \odot \mathbf{S}^5. \tag{10}$$

Note that the weight map is learned automatically during the training of the network. For more details about this weight map, please refer to [13].

The algorithm of DSiamMFT is given in Algorithm 2.

Algorithm 2: DSiamMFT

```
Input: Registered visible 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 $z_{v,1}$
Crop the infrared target image to obtain infrared template image $z_{t,1}$
Tracking:
for each frame i do
    Crop current frame images to obtain $x_v$ and $x_t$
    Compute V1 using $z_{v,1}$ and $z_{v,i-1}$
    Compute V2 using $z_{t,1}$ and $z_{t,i-1}$
    Compute W1 using $x_{v,i-1}$
    Compute W2 using $x_{t,i-1}$
    Compute $\mathbf{V1}'$, $\mathbf{V2}'$, $\mathbf{W1}'$, $\mathbf{W2}'$ accordingly
    Feed $z_v$ and $x_v$ into the visible network to obtain $f^4_{xvt}$, $f^4_{zvt}$, $f^5_{xvt}$, $f^5_{zvt}$
    Feed $z_t$ and $x_t$ into the infrared network to obtain $f^4_{xtt}$, $f^4_{ztt}$, $f^5_{xtt}$, $f^5_{ztt}$
    Fuse $f^5_{xvt}$ and $f^5_{xtt}$ to obtain fused template feature $f^5_{xvt} \oplus f^5_{xtt}$
    Fuse $f^5_{zvt}$ and $f^5_{ztt}$ to obtain fused search feature $f^5_{zvt} \oplus f^5_{ztt}$
    Compute the response map $\mathbf{S}^5$ according to Eq. (9)
    Fuse $f^4_{xvt}$ and $f^4_{xtt}$ to obtain fused search feature $f^4_{xvt} \oplus f^4_{xtt}$
    Fuse $f^4_{zvt}$ and $f^4_{ztt}$ to obtain fused template feature $f^4_{zvt} \oplus f^4_{ztt}$
    Compute the response map $\mathbf{S}^4$
    Computing the weighted-summed response map S according to
      Eq. (10)
    Upsample the response map to obtain the predicted position of
      target
end
```

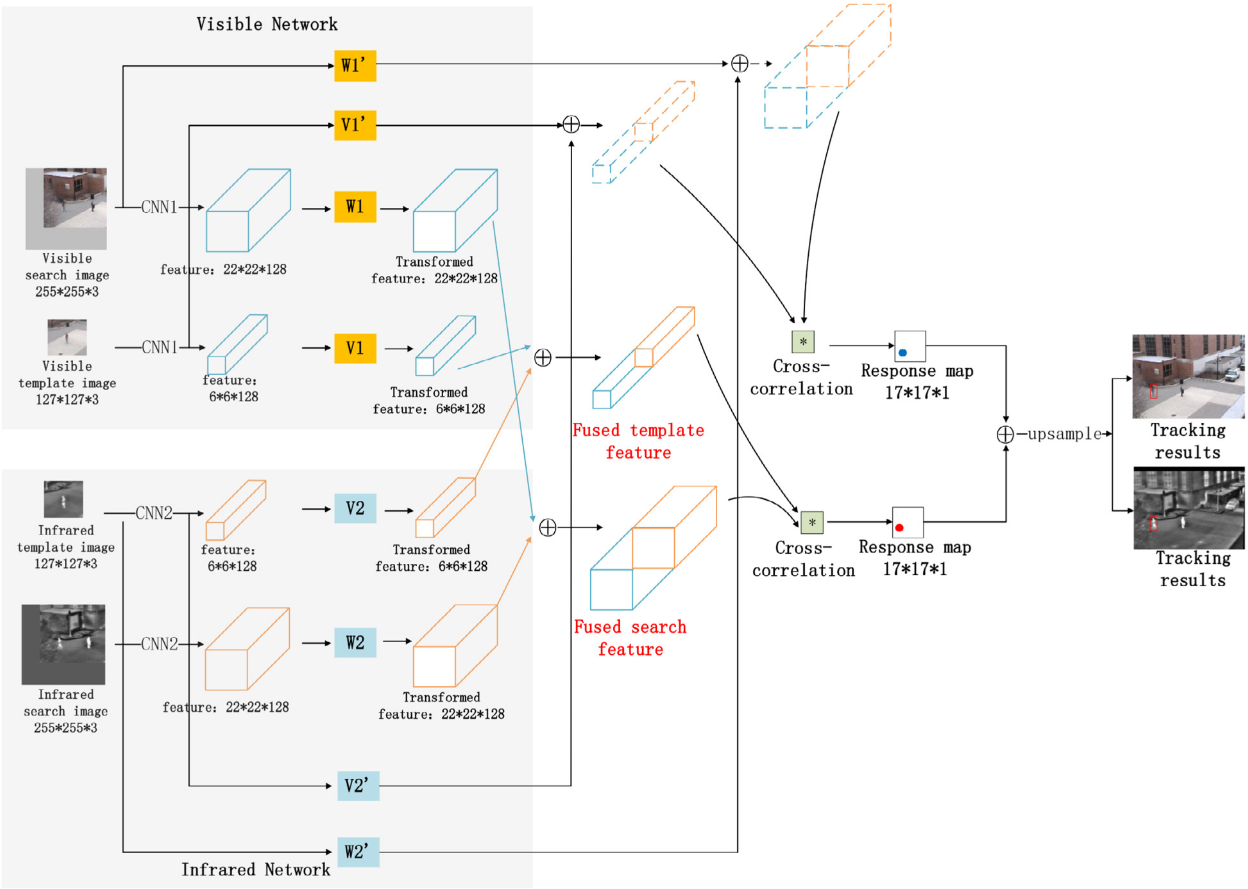


Fig. 4. Flowchart of the proposed fusion tracking method based on dynamic Siamese networks using multi-layer features.

4. Experiments

To test the performance of the proposed method, a lot of experiments are conducted. All experiments in this study are conducted using a PC with a NVIDIA GTX 1080Ti GPU and i7-8700K CPU.

4.1. Implementation details

In this work, the pretrained network provided by Guo et al. [13] is utilized in both the visible network and infrared network. The network is trained using the ImageNet dataset, which consists a large number of visible images. The main reason why we do not train both the visible and infrared networks together is that there is a lack of large-scale registered visible–infrared image pair datasets. However, although it seems a little counter-intuitive to use features trained on RGB images for thermal images, some studies have shown that the network trained using RGB images can also handle infrared images because the RGB and infrared images have something in common [36,37]. Therefore, in this study, we utilize the network provided by Guo et al. [13] as the infrared network. Besides, this study mainly aims to demonstrate the effectiveness of the dynamic Siamese networks in RGB-T fusion tracking. Therefore, we leave the training of infrared networks in our future work. We expect that by training or fine-tuning the infrared network with a large number of infrared images, the performance of the proposed RGB-T tracker could be further improved.

4.2. Datasets

In this study, a recently released large-scale RGBT dataset, namely RGBT210 is employed [22]. It contains 210 aligned visible and infrared videos. Different attributes of videos are also annotated for RGBT210 as shown in Table 1. In addition, as performed in [12,24], sixteen visible and infrared video pairs covering various challenges are also collected to evaluate the performance of the proposed method. The name and corresponding attributes of these sequences are listed in Table 2[1].

4.3. Evaluation metrics

In this study, we utilize success plot and precision plot to evaluate fusion tracking performance. Success means that the overlapping between the predicted bounding box and groundtruth box is larger than a threshold, where the overlapping is defined as:

$$O(a,b) = \frac{|a \bigcap b|}{|a \bigcup b|}, \tag{11}$$

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 smaller than a chosen threshold. The precision plot shows the trends when the threshold changes from small to large. The threshold is set to 20 pixels in this work for RGBT210. For another sixteen videos, because targets are relative small, thus the threshold is chosen as 5 pixels.

[1] SO means small object, LSV means large scale variation. Others are the same with the definition in RGBT210.

Table 2
Sequence name and attributes used in this work.

Name	Attribute	Name	Attribute
BlackCar	OCC, LSV, FM, LI	OccCar-1	OCC, LSV, TC, SO
Cycling	LSV, LI, DEF	Otvbus	LR, DEF, SO
DarkNig	LI, TC, DEF	RainyCar1	OCC, LSV, LI, TC , LR, SO
Exposure4	OCC, LSV, TC, LR, SO	RainyCar2	OCC, LSV, FM, LI, TC , LR, SO
FastMotor	OCC, LSV, TC, LR, SO	Torabi	OCC, DEF
FastMotorNig	OCC, LSV, LI, TC, LR, SO	Tricycle	OCC
GarageHover	LI, DEF	tunnel	LI, DEF
Minibus	LSV, LI	WalkingNig	OCC, LSV, LI, TC, LR, DEF, SO

Table 3
Success rate (SR %) on the RGBT210 dataset. The best three results are shown in red, green and blue, respectively. Best viewed in color.

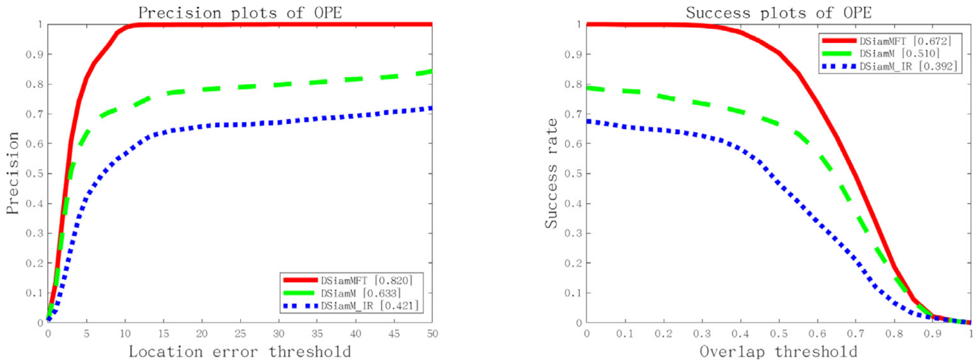
	CSR [25]	JSR [7]	MEEM+RGBT [38]	KCF+RGBT [2]	CFNet+RGBT [29]	Zhai et al. [11]	Li et al. [26]	SGT [22]	LGMG [27]	Ours
NO	45.2	27.3	41.2	36.3	52.2	46.9	59.4	50.7	58.4	56.9
PO	36.6	23.7	35.5	31.6	38.5	41.5	52.2	48.3	52.5	46.2
HO	24.3	16.5	24.2	22.2	27.3	27.0	34.6	34.1	35.9	35.2
LI	31.1	24.1	25.6	30.4	33.6	36.8	46.4	44.7	45.9	43.3
LR	23.1	15.4	23.4	26.2	27.7	33.9	37.4	37.5	35.8	31.3
TC	29.3	16.6	35.6	24.1	29.4	30.1	43.0	40.7	38.3	42.9
DEF	33.0	20.8	33.5	29.5	35.2	37.9	45.8	45.9	45.1	43.7
FM	25.0	11.9	26.8	19.1	23.0	24.9	34.9	33.1	32.5	36.7
SV	37.5	22.8	33.0	27.5	40.6	37.1	49.2	41.7	50.2	46.7
MB	23.8	14.9	31.4	20.7	22.4	25.2	40.5	39.6	42.4	38.1
CM	27.4	19.8	31.9	26.0	27.9	31.4	41.8	40.7	43.0	43.2
BC	23.7	16.9	23.4	25.6	28.1	31.7	35.2	35.5	35.2	35.0
ALL	33.0	21.3	31.9	28.5	36.0	36.6	46.3	43.0	46.8	43.6

Table 4
Success rate (SR %) on 16 sequence pairs. The best three results are shown in red, green and blue, respectively. Best viewed in color.

	CN [39]	JSR [7]	CSK [40]	CT [41]	L1 [42]	MIL [43]	RPT [44]	STC [45]	STRUCK [46]	TLD [47]	SGT [22]	LGMG [27]	Ours
BlackCar	27.1	25.5	23.2	25.0	67.8	24.1	36.9	35.0	26.7	54.6	52.8	26.9	53.3
Cycling	61.5	50.6	59.9	55.7	36.3	62.9	58.3	46.4	61.1	60.4	64.8	60.9	65.3
DarkNig	80.6	65.0	77.9	67.9	79.5	74.1	78.1	65.2	69.5	56.6	77.6	77.1	63.3
Exposure4	55.0	56.1	19.3	44.3	49.2	28.0	65.6	34.4	56.2	23.3	64.0	76.7	77.1
FastMotor	46.6	40.1	47.7	46.5	2.0	47.0	48.8	26.8	47.6	1.0	42.9	52.3	62.2
FastMotorNig	37.0	51.4	36.2	51.1	9.8	64.9	43.5	63.7	54.1	47.1	64.6	57.7	56.7
GarageHover	66.1	49.9	72.9	67.0	10.8	43.2	75.2	55.9	61.1	38.1	67.1	69.2	74.5
Minibus	42.1	41.7	41.9	42.1	37.8	34.3	43.9	46.6	43.2	42.2	74.4	45.8	76.6
OccCar-1	48.3	7.1	48.1	48.7	80.4	41.7	72.0	48.0	47.9	72.5	73.6	48.7	74.0
Otvbvs	54.0	8.8	2.3	77.3	7.8	73.2	53.0	31.0	70.8	9.7	65.8	58.8	71.1
RainyCar1	58.8	5.2	59.4	59.7	7.8	7.0	71.7	48.3	61.3	67.1	56.3	62.5	62.0
RainyCar2	56.3	52.4	39.5	39.5	56.4	45.1	59.1	47.3	56.7	47.1	64.5	55.8	75.8
Torabi	4.6	63.7	4.5	55.6	5.9	39.6	4.2	63.7	4.3	10.2	62.7	57.1	65.4
Tricycle	68.4	67.7	68.4	61.5	68.1	69.4	73.5	65.6	68.1	59.7	73.2	73.3	72.4
tunnel	81.6	31.8	74.4	61.6	79.2	57.8	76.1	60.0	73.3	68.0	77.3	74.1	75.7
WalkingNig	63.1	43.8	63.6	16.6	50.7	20.6	52.9	61.4	56.7	16.3	61.4	60.7	52.6
ALL	53.2	41.3	44.2	51.2	40.9	45.8	57.1	50.0	53.7	42.1	59.0	66.1	67.2

Table 5
Precision rate (PR %) on 16 sequence pairs. The best three results are shown in red, green and blue, respectively. Best viewed in color.

	CN [39]	JSR [7]	CSK [40]	CT [41]	L1 [42]	MIL [43]	RPT [44]	STC [45]	STRUCK [46]	TLD [47]	SGT [22]	LGMG [27]	Ours
BlackCar	16.5	11.3	7.0	3.5	80.9	3.5	12.2	10.4	14.8	46.1	14.8	28.7	56.5
Cycling	82.5	41.9	72.5	27.5	45.6	88.1	58.8	99.4	48.1	61.9	79.4	85.0	39.4
DarkNig	100	68.5	99.1	59.5	98.2	79.3	100	100	56.8	87.4	87.4	98.2	38.7
Exposure4	75.5	86.4	26.5	48.3	72.1	39.5	98.6	53.1	81.6	25.2	88.4	98.0	95.2
FastMotor	40.0	18.0	100	90.0	2.0	77.0	76.0	19.0	47.0	1.0	100	100	100
FastMotorNig	30.8	60.0	44.6	64.6	13.8	90.8	58.5	100	69.2	64.2	96.9	100	83.1
GarageHover	88.0	44.6	98.8	59.4	13.1	11.2	100	58.2	72.9	21.1	96.0	94.4	99.6
Minibus	54.7	20.5	62.1	17.9	68.4	18.9	65.3	79.5	78.4	47.4	100	100	100
OccCar-1	95.9	7.6	83.6	79.5	98.2	16.4	98.2	91.8	42.7	86.5	98.3	98.3	100
Otvbvs	54.1	10.5	2.6	98.3	10.0	94.9	37.9	13.7	99.1	13.4	100	97.2	100
RainyCar1	100	5.0	100	96.7	8.3	8.3	96.7	5.0	100	85.0	100	98.3	96.7
RainyCar2	76.8	68.0	49.6	43.2	74.4	40.8	75.2	74.4	73.6	50.4	74.4	82.4	96.0
Torabi	1.7	32.1	1.7	7.5	2.5	2.9	1.7	40.8	1.7	2.9	10.0	11.3	19.9
Tricycle	95.4	97.7	98.5	79.2	73.1	95.4	100	75.4	100	89.2	99.2	99.2	100
tunnel	100	20.0	89.0	12.0	100	50.0	95.0.	60.5	98.0	42.5	98.5	98.0	98.5
WalkingNig	100	46.7	81.4	18.6	100	25.7	97.6	100	95.8	15.6	97.1	96.4	88.0
ALL	69.5	39.9	61.4	50.3	53.8	46.4	73.2	61.3	67.5	46.3	83.8	86.6	82.0



(a) Precision rate

(b) Success rate

Fig. 5. The results comparison of DSiamM, DSiamM_IR and DSiamMFT on sixteen sequence pairs.

5. Results

5.1. Evaluation on RGBT210 dataset

The comparison of tracking performance between DSiamMFT and several state-of-the-art tracking algorithms, including CSR [25], JSR [7], KCF [2], CFNet [29], SGT [22], LGMG [27], MEEM [38], the method proposed by Zhai et al. [11] and the method of Li et al. [26], on RGBT210 dataset is given in Table 3. The table shows that DSiamMFT obtains very competitive results with the state-of-the-arts, ranking the third in terms of SR. In particular, DSiamMFT outperforms the famous KCF over 15.1% and the correlation filter-based fusion trackers recently proposed by Zhai et al. [11] over 7.7% in SR, respectively. Besides, although DSiamMFT performs slightly worse than the method proposed by Li et al. [26] and LGMG, it beats all trackers in several challenging scenarios, namely fast motion and camera moving. DSiamMFT also ranks the second in handling heavy occlusion and thermal crossover. This clearly demonstrate that the proposed DSiamMFT can effectively handle these adverse challenging conditions during tracking by leveraging complementary multi-modal information. Besides, as will be discuss later, the proposed DSiamMFT runs much faster (14.7 FPS) than the method of Li et al. [26] (8 FPS) and LGMG (7 FPS), thus striking a better balance between tracking precision and speed.

5.2. Evaluation on 16 sequences

The performance of DSiamMFT on sixteen sequences are compared with several state-of-the-art trackers, including SGT [22], LGMG [27], CN [39], JSR [7], CSK [40], CT [41], L1 [42], MIL [43], RPT [44], STC [45], STRUCK [46], TLD [47]. The results are shown in Table 4 and Table 5. As can be seen, the proposed DSiamMFT achieves the best result in SR and the third best result in PR, demonstrating the effectiveness of DSiamMFT in RGB-T fusion tracking. In particular, DSiamMFT outperforms all other compared trackers in 7 sequences in terms of SR and in 6 sequences in terms of PR. These sequences cover all annotated attributes namely OCC, LSV, LI, TC, LR, DEF, SO. This further shows that the proposed DSiamMFT is effective in handling various challenging scenarios during tracking.

5.3. Ablation study

There are several important components in the proposed method, namely the feature-level fusion, the multi-modal images and the multi-layer fusion. To investigate the contribution of each component to the tracking results, a series of ablation studies have been performed.

Table 6
Image fusion methods employed in this study.

Name	Description	Method denotation
Fused 1	Average	DSiamM_Average
Fused 2	TGB	DSiamM_TGB
Fused 3	RTB	DSiamM_RTB
Fused 4	RGT	DSiamM_RGT
Fused 5	ADF	DSiamM_ADF

5.3.1. Contribution of feature-level fusion

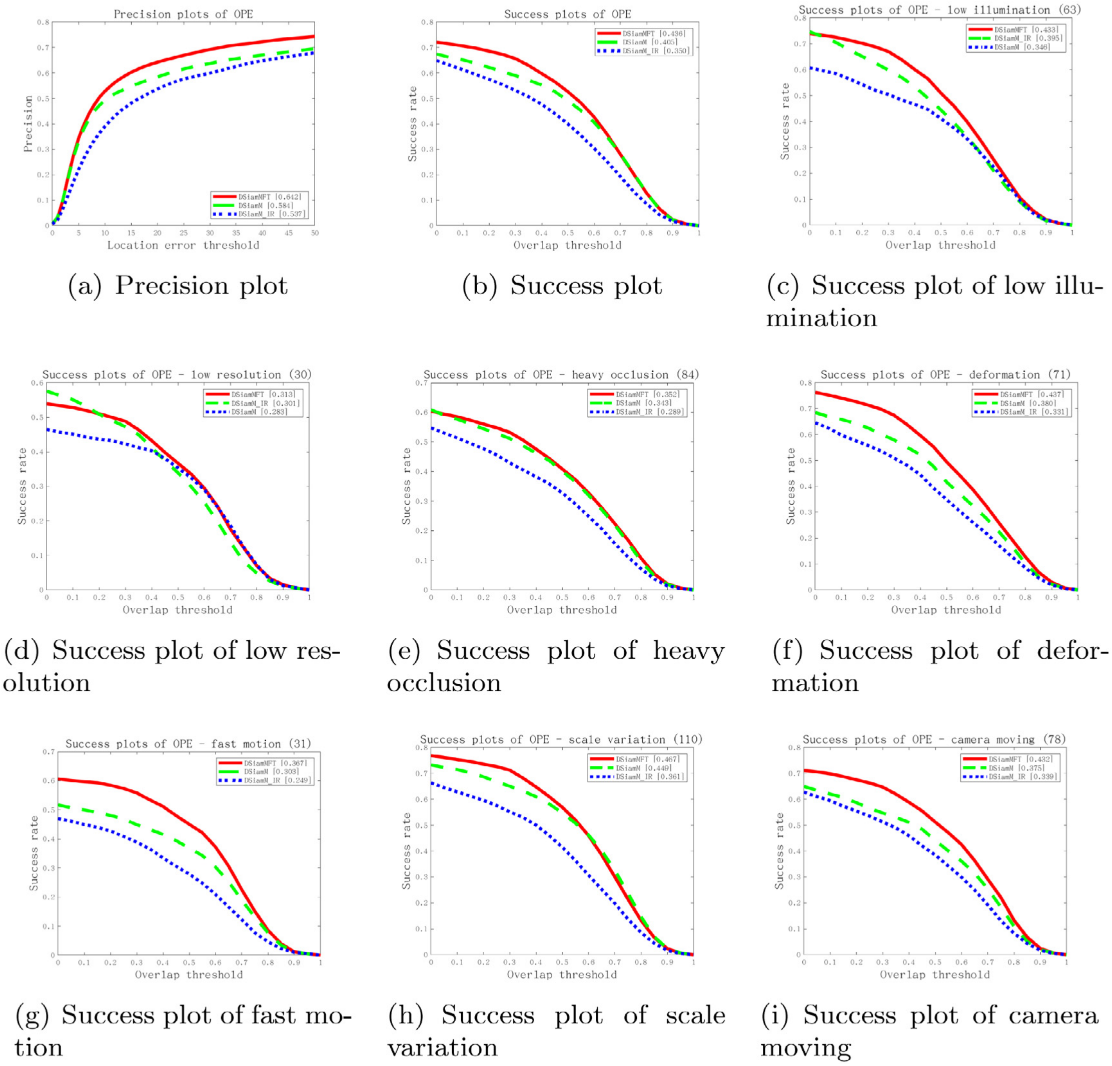
To demonstrate the contribution of feature-level fusion utilized in this work, we compare the results of the proposed method with baselines which simply perform early fusion of two modalities as input to a single dynamic Siamese network [13]. Similar to the work of Zhang et al. [48], we chose five image fusion methods to firstly produce five fused version of RGBT210 dataset. Then, we input the fused image to DSiamM [13] to obtain tracking results. The five fusion methods are chosen as listed in Table 6. In the first method, we simply compute the average between visible and infrared images as

$$I_f = 0.5 \times I_v + 0.5 \times I_i, \tag{12}$$

where I_f denotes fused images, I_v and I_i are the visible and infrared images, respectively. From Fused 2 to Fused 4 methods, we replace one channel in visible images using the corresponding infrared image, resulting in TGB, RTB and RGT images. The fifth method is called ADF [49]. Default settings proposed by the corresponding authors are chosen.

The results comparison of the proposed method with those early fusion methods is presented in Table 7. It can be seen that the proposed DSiamMFT obtains the best performance in terms of both PR and SR than its early fusion counterparts. This clearly demonstrate the superiority of the proposed DSiamMFT.

Table 7 shows that the speed of DSiamMFFT is much slower than those pixel-level fusion tracking algorithms. However, it should be mentioned that, the speed listed in Table 7 is the speed of running DSiamM on each fused dataset which does not contain the time of performing image fusion. Actually, the first four pixel-level fusion methods are quick, while the ADF method is relatively time-consuming. It takes around 0.66 s to fuse one visible and infrared image pair [49]. Indeed, if we do both image fusion and tracking online, the image fusion speed will significantly affect the whole fusion tracking speed. If an image fusion algorithm is very slow, for instance it takes about 80 s to fuse one image pair using the latent low-rank representation (LatLRR) [50], then it will not be feasible to obtain a real-time fusion tracking approach.



(a) Precision plot

(b) Success plot

(c) Success plot of low illumination

(d) Success plot of low resolution

(e) Success plot of heavy occlusion

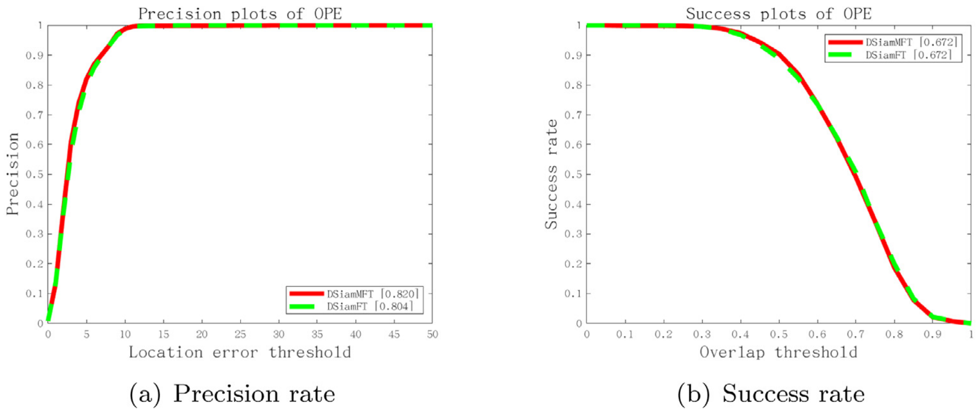
(f) Success plot of deformation

(g) Success plot of fast motion

(h) Success plot of scale variation

(i) Success plot of camera moving

Fig. 6. The results comparison on RGBT210. The title of each plot is the attribute name and corresponding sequences number.



(a) Precision rate

(b) Success rate

Fig. 7. The results comparison of DSiamFT and DSiamMFT on sixteen sequence pairs.

Table 7
Comparison of results (PR %, SR% and FPS) between the proposed method and early fusion algorithms. The best three results are shown in red, green and blue, respectively. Best viewed in color.

Att	DSiamM_Average	DSiamM_TGB	DSiamM_RTB	DSiamM_RGT	DSiamM_ADF	DSiamMFT
PR	55.4	55.3	52.2	58.6	56.2	64.2
SR	38.2	37.6	34.2	40.9	38.1	43.6
Speed	33.9	31.0	30.5	30.3	33.2	14.7

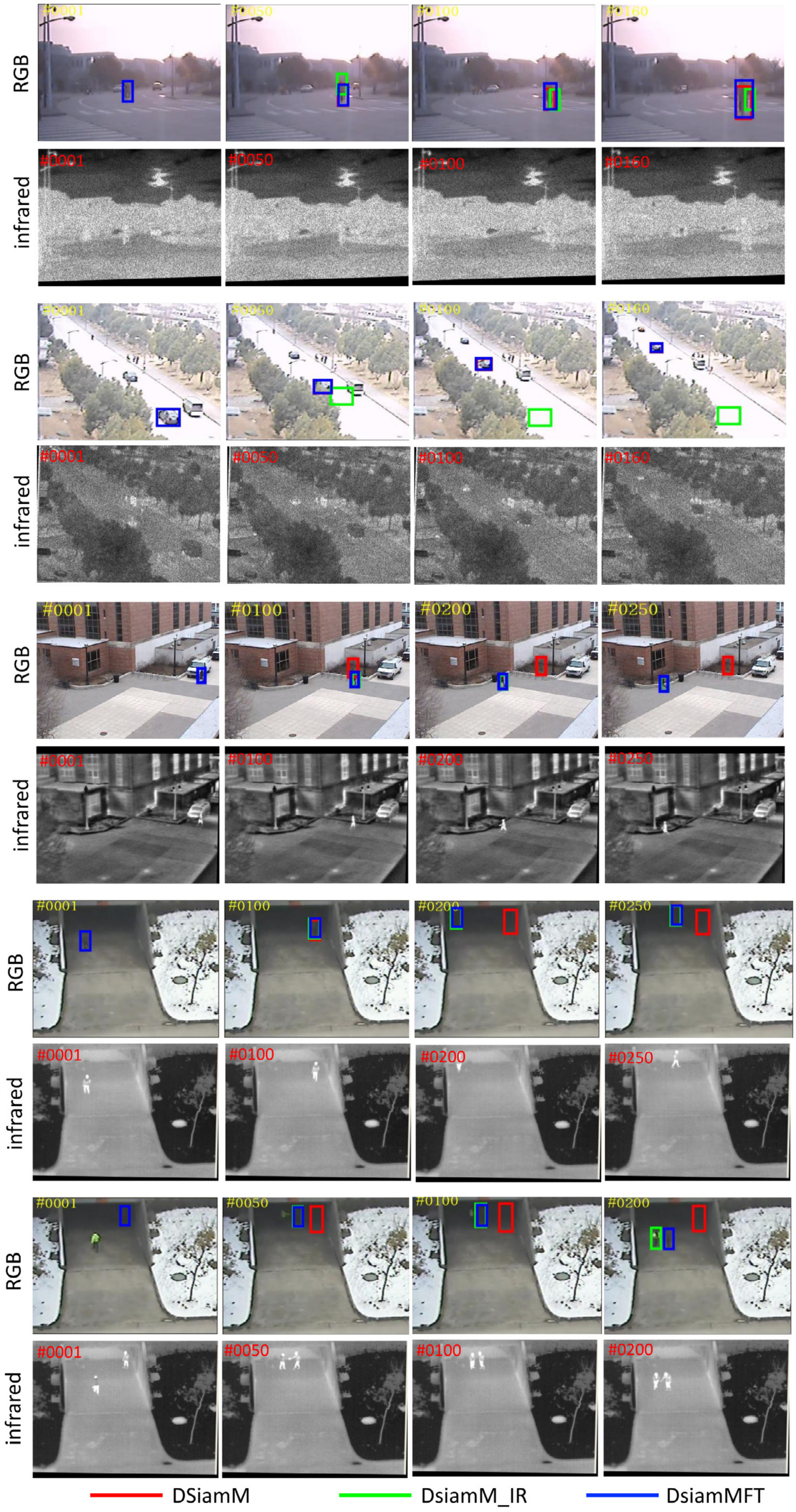


Fig. 8. Qualitative comparison between DSiamMFT, DSiamM and DSiamM_IR. From top to bottom are sequences: *Cycling*, *OccCar-1*, *otvbvs*, *GarageHover*, *tunnel*. In *Cycling* and *OccCar-1*, DSiamM_IR cannot track the target well. In *otvbvs* and *GarageHover*, DSiamM loses the target. In *tunnel*, both DSiamM and DSiamM_IR fail. In all these sequences, the proposed DSiamMFT can track the target successfully.

Table 8
Comparison of running time (FPS) between the proposed method and other reported RGB-T trackers. The best three results are shown in red, green and blue, respectively. Best viewed in color.

RPT [44]	Struck [46]	CSR [25]	SGT [22]	Li et al. [26]	Zhai et al. [11]	LGMG [27]	DSiamFT	DSiamMFT
2.6	10.8	1.6	5	8	227	7	17	14.7

5.3.2. Contribution of multi-modal images

To demonstrate the contribution of multi-modal information fusion during tracking, we have run tracking with single modality images. We firstly run the tracker on sixteen sequences with only visible images and only infrared sequences, and term them as DSiamM and DSiamM_IR, respectively. The comparison of results between DSiamM, DSiamM_IR and DSiamMFT on these sequences is shown in Fig. 5. It can be observed clearly that DSiamMFT outperforms both DSiamM and DSiamM_IR with a very clear margin, which means that by using multi-modal information, the tracking performances have been improved significantly.

We then run DSiamM and DSiamM_IR on RGBT210 dataset, and the results are shown in Fig. 6. As can be seen, in all these cases DSiamMFT outperforms both DSiamM and DSiamM_IR with a very clear margin, indicating that by utilizing complementary information from visible and infrared images, the tracking performance can be improved greatly in various adverse challenging situations. Besides, in all presented attributes except for low illumination and low resolution, DSiamM outperforms DSiamM_IR. This indicates that tracking based on visible images may not work well when illumination condition is poor and when the resolution of images is low. Thus, visible and infrared images indeed contribute complementary features in fusion tracking.

5.3.3. Contribution of multi-layer fusion

To investigate the effectiveness of multi-layer fusion in the tracking process, we compare the performance of DSiamFT and DSiamMFT. The comparison of PR and SR on sixteen sequences are given in Fig. 7. As can be seen, after removing the multi-layer fusion in the fusion tracking process, the tracking precision degrades. This clearly demonstrates that the multi-layer fusion contributes to the performance of DSiamMFT.

5.4. Qualitative tracking results

To further illustrate the performance of the proposed multi-modal fusion tracking algorithm, some qualitative tracking results are presented in Fig. 8. Both visible and infrared images of five sequences are shown for better visualization. In sequence *Cycling* and *OccCar-1*, both DSiamM and DSiamMFT track the target successfully, while DSiamM_IR fail. The possible reasons are the thermal crossover and low resolution of infrared images. In contrast, in sequence *otvbvs* and *GarageHover*, DSiamMFT and DSiamM_IR track the object well, while DSiamM loses the target. In *otvbvs*, the black pillar is distraction to the man in black cloth, thus DSiamM shifts to the pillar and fails. In this case, the infrared image reveal different thermal information between the man and the pillar, thus DSiamM_IR can track successfully. In *GarageHover*, the illumination condition around the target is very poor thus DSiamM is not able to track it. In the sequence *tunnel*, again the illumination condition around the target is poor. Besides, another person which has similar thermal information with the target appears and is a distraction to DSiamM_IR. Therefore, in this case both DSiamM and DSiamM_IR fail.

In all these five sequences, DSiamMFT tracks the target successfully, demonstrating that DSiamMFT is able to handle challenging scenarios such as poor illumination, background clutter and thermal crossover. Furthermore, the results on sequence *tunnel* demonstrate that the proposed DSiamMFT can even work in some cases where both DSiamM and DSiamM_IR fail. This clearly shows the benefits of fusing complementary features from visible and infrared images in tracking.

5.5. Runtime performance

The comparison of runtime between the proposed fusion tracking method with several state-of-the-arts trackers is given in Table 8. As can be seen, DSiamFT can run almost in real time with a framerate of 17 FPS. DSiamMFT is slightly slower but is still faster than all trackers except for the tracker proposed by Zhai et al. [11] which is based on correlation filter. However, Table 3 indicates that DSiamMFT outperforms their tracker on RGBT210 datasets with a clear margin in terms of both precision rate and success rate. In addition, although SGT [22] and the methods proposed by Li et al. [26] obtain competitive or slightly better results compared with DSiamMFT in terms of PR and SR, these trackers are very slow and their frame rate are only 5 FPS and 8 FPS, respectively. This means that they are not suitable to be used in practical real-time applications. Therefore, the proposed DSiamMFT strikes a better balance between tracking precision and speed than these compared trackers.

6. Discussion

Improvements in almost all attribute-based performance. The original purpose of combining infrared images in tracking is to improve performance when visible images are unreliable. For example, when light conditions are poor. However, experimental results are supervising, since the fusion tracking performance of almost all attributes on RGBT210 have been improved compared to those of visible images. This indicates clearly that infrared images are very beneficial in object tracking. We believe that this is because infrared images can provide thermal features which are more robust in some challenging situations, and the proposed fusion tracking method can effectively make use of complementary features from both modalities.

Infrared-specific network. In this work, the network pretrained with ImageNet are utilized directly without finetuning as a proof of concept. Note that the ImageNet dataset does not contain infrared images. Despite these factors, the proposed DSiamMFT still achieves very competitive results again the start-of-the-arts. This clearly demonstrate the strength of the proposed method. We believe that by fine-tuning the infrared network with infrared images, or by training an infrared network from scratch using infrared images may further improve the performance of DSiamMFT.

7. Conclusion

In this paper, a fusion tracking method based on visible and infrared thermal images (RGB-T) via dynamic Siamese Networks with multi-layer fusion, termed as DSiamMFT, is proposed. To the best of our knowledge, this is the first time that the dynamic Siamese network is leveraged to perform RGB-T fusion tracking. Specifically, two dynamic Siamese networks, namely visible network and infrared network, are employed to process visible and infrared images, respectively. Features of the last layer in visible network are fused with those in infrared network. Similarly, features of the last second layer in visible network are fused with those in infrared network. Response maps produced through cross-correlation using different fused layer features are integrated to produce the final response map which can be used to predict target location. Extensive experiments have been conducted, which clearly indicate that the proposed DSiamMFT can improve tracking performance significantly compared to tracking with single-modal images. Also, the proposed network architecture is simple yet effective, thus it can run at a faster speed than most RGB-T trackers while achieving competitive performance against the state-of-the-arts although without fine-tuning.

Acknowledgment

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. The authors would like to thank Prof. Chenglong Li from Anhui University for providing some benchmark results.

References

[1] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H. Torr, Fully-convolutional siamese networks for object tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 850–865.

[2] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 583–596.

[3] P. Li, D. Wang, L. Wang, H. Lu, Deep visual tracking: Review and experimental comparison, Pattern Recognit. 76 (2018) 323–338.

[4] J.W. Davis, V. Sharma, Fusion-based background-subtraction using contour saliency, in: Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2005, p. 11.

[5] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, G. Xiao, SiamFT: An RGB-infrared fusion tracking method via fully convolutional siamese networks, IEEE Access 7 (2019) 122122–122133.

[6] C. Li, X. Liang, Y. Lu, N. Zhao, J. Tang, RGB-T object tracking: benchmark and baseline, 2018, arXiv preprint arXiv:1805.08982.

[7] H. Liu, F. Sun, Fusion tracking in color and infrared images using joint sparse representation, Sci. China Inf. Sci. 55 (3) (2012) 590–599.

[8] C. Li, X. Wu, N. Zhao, X. Cao, J. Tang, Fusing two-stream convolutional neural networks for RGB-T object tracking, Neurocomputing 281 (2018) 78–85.

[9] Y. Wang, C. Li, J. Tang, Learning soft-consistent correlation filters for RGB-T object tracking, in: Chinese Conference on Pattern Recognition and Computer Vision, PRCV, Springer, 2018, pp. 295–306.

[10] X. Gang, Y. Xiao, J. Wu, A new tracking approach for visible and infrared sequences based on tracking-before-fusion, Int. J. Dyn. Control 4 (1) (2016) 40–51.

[11] S. Zhai, P. Shao, X. Liang, X. Wang, Fast RGB-T tracking via cross-modal correlation filters, Neurocomputing 334 (2019) 172–181.

[12] X. Lan, M. Ye, R. Shao, B. Zhong, P.C. Yuen, H. Zhou, Learning modality-consistency feature templates: A robust RGB-infrared tracking system, IEEE Trans. Ind. Electron. 66 (12) (2019) 9887–9897.

[13] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic siamese network for visual object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1763–1771.

[14] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: Recent advances and future prospects, Inf. Fusion 42 (2018) 158–173.

[15] X. Yan, S.Z. Gilani, H. Qin, A. Mian, S. Member, S.Z. Gilani, H. Qin, A. Mian, Unsupervised deep multi-focus image fusion, 2018, pp. 1–11, arXiv preprint arXiv:1806.07272.

[16] K. Xia, H. Yin, J. Wang, A novel improved deep convolutional neural network model for medical image fusion, Cluster Comput. 22 (2018) 1515–1527.

[17] H. Li, X. Wu, DenseFuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.

[18] K.R. Prabhakar, V.S. Srikar, R.V. Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017, pp. 4724–4732.

[19] H. Hermessi, O. Mourali, E. Zagrouba, Convolutional neural network-based multimodal image fusion via similarity learning in the shearlet domain, Neural Comput. Appl. (2018) 1–17.

[20] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (June 2018) (2019) 11–26.

[21] Y. Liu, X. Chen, J. Cheng, H. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, Int. J. Wavelets Multiresolut. Inf. Process. 16 (03) (2018) 1850018.

[22] C. Li, N. Zhao, Y. Lu, C. Zhu, J. Tang, Weighted sparse representation regularized graph learning for RGB-T object tracking, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017, pp. 1856–1864.

[23] X. Lan, M. Ye, S. Zhang, P.C. Yuen, Robust collaborative discriminative learning for RGB-infrared tracking, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7008–7015.

[24] X. Lan, M. Ye, S. Zhang, H. Zhou, P.C. Yuen, Modality-correlation-aware sparse representation for RGB-infrared object tracking, Pattern Recognit. Lett. 130 (2018) 12–20.

[25] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, L. Lin, Learning collaborative sparse representation for grayscale-thermal tracking, IEEE Trans. Image Process. 25 (12) (2016) 5743–5756.

[26] C. Li, C. Zhu, Y. Huang, J. Tang, L. Wang, Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking, in: Proceedings of ECCV, 2018, pp. 808–823.

[27] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, J. Tang, Learning local-global multi-graph descriptors for RGB-T object tracking, IEEE Trans. Circuits Syst. Video Technol. 29 (10) (2018) 2913–2926.

[28] N. Xu, G. Xiao, X. Zhang, D.P. Bavirisetti, Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences, in: Proceedings of the 4th International Conference on Virtual Reality, ACM, 2018, pp. 44–49.

[29] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H. Torr, End-to-end representation learning for correlation filter based tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2805–2813.

[30] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4834–4843.

[31] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 101–117.

[32] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.

[33] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, SiamRPN++: Evolution of siamese visual tracking with very deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4282–4291.

[34] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond, MIT press, 2001.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of Computer Vision and Pattern Recognition, CVPR. IEEE Conference on, IEEE, 2009, pp. 248–255.

[36] Q. Liu, X. Lu, Z. He, C. Zhang, W.-S. Chen, Deep convolutional neural networks for thermal infrared object tracking, Knowl.-Based Syst. 134 (2017) 189–198.

[37] X. Li, Q. Liu, N. Fan, Z. He, H. Wang, Hierarchical spatial-aware siamese network for thermal infrared object tracking, Knowl.-Based Syst. 166 (2019) 71–81.

[38] J. Zhang, S. Ma, S. Sclaroff, MEEM: robust tracking via multiple experts using entropy minimization, in: European Conference on Computer Vision, Springer, 2014, pp. 188–203.

[39] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1090–1097.

[40] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: European Conference on Computer Vision, Springer, 2012, pp. 702–715.

[41] D. Wang, H. Lu, Visual tracking via probability continuous outlier model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3478–3485.

[42] Y. Wu, E. Blasch, G. Chen, L. Bai, H. Ling, Multiple source data fusion via sparse representation for robust visual tracking, in: 14th International Conference on Information Fusion, IEEE, 2011, pp. 1–8.

[43] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1619–1632.

[44] Y. Li, J. Zhu, S.C. Hoi, Reliable patch trackers: Robust visual tracking by exploiting reliable patches, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 353–361.

[45] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1409–1422.

[46] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S.L. Hicks, P.H. Torr, Struck: Structured output tracking with kernels, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2016) 2096–2109.

[47] K. Zhang, L. Zhang, M.-H. Yang, Real-time compressive tracking, in: European Conference on Computer Vision, Springer, 2012, pp. 864–877.

[48] X. Zhang, G. Xiao, P. Ye, D. Qiao, J. Zhao, S. Peng, Object fusion tracking based on visible and infrared images using fully convolutional siamese networks, in: Proceedings of the 22nd International Conference on Information Fusion, IEEE, 2019.

[49] D.P. Bavirisetti, R. Dhuli, Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform, IEEE Sens. J. 16 (1) (2016) 203–209.

[50] H. Li, X. Wu, Infrared and visible image fusion using latent low-rank representation, 2018, arXiv preprint arXiv:1804.08992.