Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07867-1

ORIGINAL ARTICLE

Robust thermal infrared tracking via an adaptively multi-feature fusion model

Di Yuan[1] · Xiu Shu[2] · Qiao Liu[3] · Xinming Zhang[2] · Zhenyu He[4]

Received: 25 April 2022 / Accepted: 21 September 2022

Abstract
When dealing with complex thermal infrared (TIR) tracking scenarios, the single category feature is not sufficient to portray the appearance of the target, which drastically affects the accuracy of the TIR target tracking method. In order to address these problems, we propose an adaptively multi-feature fusion model (AMFT) for the TIR tracking task. Specifically, our AMFT tracking method adaptively integrates hand-crafted features and deep convolutional neural network (CNN) features. In order to accurately locate the target position, it takes advantage of the complementarity between different features. Additionally, the model is updated using a simple but effective model update strategy to adapt to changes in the target during tracking. In addition, a simple but effective model update strategy is adopted to adapt the model to the changes of the target during the tracking process. We have shown through ablation studies that the adaptively multi-feature fusion model in our AMFT tracking method is very effective. Our AMFT tracker performs favorably on PTB-TIR and LSOTB-TIR benchmarks compared with state-of-the-art trackers.

Keywords Thermal infrared tracking · Multi-feature fusion · Model update

1 Introduction

Researchers are becoming increasingly interested in TIR target tracking due to its effectiveness in dark environments [1–4]. Although the current research has made some progress, the tracking result is not ideal due to the lack of a single feature's ability to express the target is still a problem worthy of research [5–7].

Recently, influenced by the success of the CNN architecture in various visual tasks [8–16], some methods have attempted to use the CNN's powerful representation capabilities to improve the TIR target tracking performance [2, 17–20]. The MCFTS [2] tracker uses the pre-trained CNN to extract multi-layer convolution features of the infrared target and combines the kernel correlation filters method to construct an integrated TIR tracking method, which has achieved good tracking results. Gao et al. [17] introduced the pre-trained deep appearance features and deep optical flow features into the structure output support vector machine for TIR target tracking. Li et al. [21] proposed a TIR tracking method using sparse representation of deep semantic features. The deep semantic feature is obtained through a pre-trained convolutional neural network combined with a target feature channel selection module based on a supervised training method. The HSSNet [20] tracker trains a hierarchical spatial perception feature model end-to-end under the framework of the matching task to represent the TIR image target and designs a matching TIR tracking method. To adapt to the change of target appearance, Gundogdu et al. [4] proposed an integrated TIR tracking method based on correlation filters using convolutional neural networks features. Zhang et al. [19] proposed to use the generative confrontation network to convert visible light images into the TIR images and use these synthesized the TIR images to train a twin

✉ Di Yuan
yuandi@xidian.edu.cn

1 Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China

2 School of Science, Harbin Institute of Technology, Shenzhen 518055, China

3 National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China

4 School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China

Published online: 12 October 2022

network based on matching. Then use the twin network to extract the deep features of the TIR targets and integrate these features for TIR target tracking. Although some progress has been made with the above-mentioned methods, since a single type of feature cannot fully characterize the appearance information of the target, the characteristics of the TIR target obtained based on these networks can not achieve the optimal tracking result for the target tracking.

To solve all of these issues, we developed an adaptively multi-feature fusion model (AMFT) capable of tracking TIR targets with high efficiency and robustness. Generally speaking, it can be said that hand-crafted features have good spatial structure information, which makes it easier to distinguish targets from backgrounds, but their ability to characterize the target is obviously insufficient. While the deep convolutional neural networks features with discriminative semantic information can help us accurately detect the position of the target, they cannot adapt to changes in the spatial location of the target. Our AMFT tracking method can adaptively fuse the advantages of hand-crafted features and deep convolutional neural networks features so that it can track targets more accurately and robustly. Figure 1 shows that the tracking results of our AMFT tracking method are similar to the ground-truth labels of the tracking target, demonstrating the effectiveness of the proposed multi-feature fusion model.

Following is a summary of the main contributions:

- An improved tracking method based on multi-feature fusion (AMFT) is proposed for the TIR target tracking task.
- The presented AMFT tracker could train a multi-feature fusion model that may autonomously integrate the direct benefits of several features to better characterize the target appearance.
- Extensive comparative evaluations show that the proposed AMFT tracker outperforms other trackers on PTB-TIR [22] and LSOTB-TIR [23] benchmarks.

2 Related works

we will actually mainly introduce some of the most relevant studies to our tracking method, including such tracking methods [24–28] and multiple features fused methods [29–34] in this section.

The correlation filters (CF) can determine the degree of similarity between the signals by performing correlation operations on two signals. For the target tracking task, it can be regarded as a similarity measurement between tracking target and candidates, and the candidate sample who with the greatest similarity to the target will be found in the search area as the tracking target. Because CF could perform a fast operation on numerous training samples, these CF-based trackers have achieved better tracking results [35–38]. Most trackers based on the CF framework use the cyclic structure of training samples to learn linear

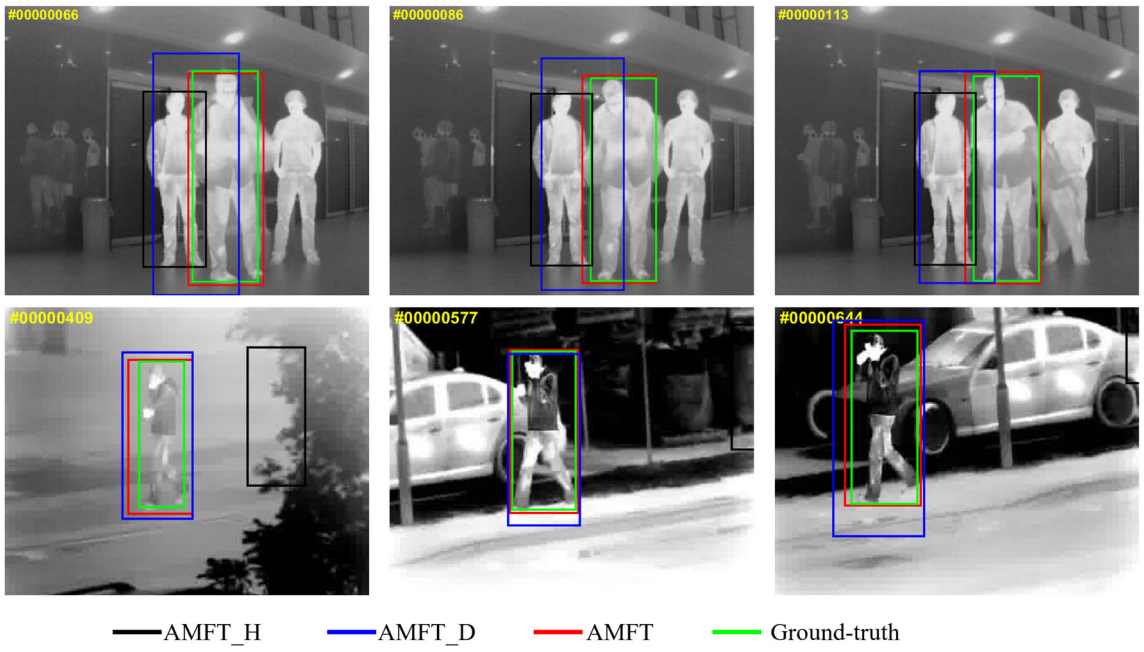

Fig. 1 Tracking examples. AMFT_H represents the tracking results with only hand-crafted features, while AMFT_D represents the tracking results with only deep convolutional neural networks features

filters. The image patch produced by the cyclic shift is similar to the translation of the target and cannot simulate the real tracking scene. When the tracking scene is more complicated, the tracking results usually obtained by relying on the response map will be inaccurate, and the target will be lost. In order to obtain the desired output response map, Bibi et al. [39] used the score of the real sample to replace the score of the cyclic shift sample, which made up for the shortcomings of the manually set response map. Based on the good properties of the CF-based tracking framework, many attempts are made to introduce it into the TIR target tracking task [1, 4, 6, 29, 40]. He et al. [1] introduce a weighted correlation filter-based infrared target tracking method to obtain efficient tracking results. Gundogdu et al. [4] verifies that good TIR target tracking accuracy can be achieved by using deep convolutional feature in the CF-based tracker.

Recently, deep learning-based trackers have achieved good results on target tracking task. Convolutional neural networks have becomes more popular in target tracking tasks due to their formidable feature extraction capabilities [41–47]. In [42], Wang et al. propose that an unsupervised CNN model be trained on large-scale unlabeled sample images, which can successfully solve the issue of insufficient training samples with labels. The Siamese network-based tracking framework approaches treat the tracking task as a template-matching problem, and returning the most similar target candidate as the tracking result by calculating the similarity between the template target and the target candidates. The SiamFC [43] tracker introduces a fully-convolutional Siamese network for the tracking task. The CFNet [44] tracker attempts to treat the correlation filters as a network layer in the deep network architecture to obtain a faster tracking speed. Dong et al. [45] introduces a triplet loss to extract expressive deep features for visual tracking tasks by adding them into the Siamese network framework instead of pairwise loss for model training. In [48], a structured target-aware model has been proposed to improve the target tracking performance in the TIR scenarios.

Multi-feature fusion is a common method to improve tracker performance in target tracking task [31–33, 49–52]. Liu et al. [31] propose to simultaneously learn local structural features and global semantic features of the TIR images under the framework of matching network to enhance the discrimination ability of feature model to similar interferers. The HDT [32] tracker integrates multiple weak correlation filters based on deep features through a Hedge method, which can be used to automatically update the weight of each weak tracker so as to locate the target more accurately. The MFFT [33] tracker adopts the complementarity between multiple different features to enhance the robustness of the proposed tracking method. In [2], a MCFTS tracker has been proposed to uses a Kullback–Leibler divergence fusion method to integrate multiple convolution feature-based correlation filters for the TIR target tracking task. Zhang et al. [53] propose a tracking framework to integrate the RGB and TIR images in the RGBT tracking task in an end-to-end way. In a deep RGBT tracking framework, Li et al. [54] describe a multi-adapter convolutional network that performs modality-shared, modality-specific, and instance-aware feature learning simultaneously.

Although these trackers have produced some acceptable tracking performance, the existing tracking approaches are still unable to obtain optimal tracking results when faced with identical object interference, occlusions, and other difficult challenges due to the complexity of the TIR tracking scenarios. We present a multi-feature fusion-based tracking approach that relies on the complementarity between different types of features to increase the tracker's tracking performance in these TIR tracking scenarios.

3 The proposed AMFT tracker

For more accurate TIR target tracking performance, we propose a multi-feature fusion model for characterizing the target appearance more comprehensively. We use the correlation filters-based tracking framework to generate the corresponding response map complementary fusion of different features to a better accurate target location. First, we briefly introduce the correlation filters-based tracking framework. After that, we propose a multi-feature fusion mode for accurate TIR target tracking results.

3.1 Correlation filters-based tracking framework

The correlation filters (CF)-based trackers have been extensively studied in recent years, and it has greatly improved the tracking speed under the premise of ensuring tracking accuracy. The CF-based tracking methods usually train a classifier to identify the target from the background [35, 36, 46]. We construct a weak target tracker using the correlation filters for every single category of features and then construct a multi-feature fusion tracker through the fusion of the response maps of multiple weak trackers to better handle the challenging problems in the TIR target tracking task. The correlation filters w_k corresponding to the k-th features can be obtained as follow:

$$w_k = \arg\min_{w_k}(||w_k * x_k - y||^2 + \lambda||w_k||^2), \quad (1)$$

where w_k is the trained correlation filters for the k-th features, x_k denotes the k-th features of the training samples, y denotes the Gaussian-shape label of training samples, and λ

is a regularization parameter. Define $\mathscr{F}$ as the Fourier transform, and Eq. (1) in the Fourier domain yields a closed solution of the following form:

$$w_k = \mathscr{F}^{-1}\left(\frac{\hat{x_k} \odot \hat{y}}{\hat{x_k}^* \odot \hat{x_k} + \lambda}\right). \tag{2}$$

where $\mathscr{F}(x_k) = \hat{x_k}$, $\mathscr{F}(y) = \hat{y}$ and $\hat{x_k}^*$ is the conjugate transpose of $\hat{x_k}$.

The main purpose of the search phase is to obtain the response map of the target position in the search image frame. First, given the search area z of a TIR image target and extract the different types of features z_k. Then, the features z_k are transformed into the Fourier domain: $\mathscr{F}(z_k) = \hat{z_k}$. Therefore, the response map of the target location dependent on the k-th features z_k can be obtained by the following cross-correlation operation:

$$R_k = \mathscr{F}^{-1}(\hat{z_k} \otimes \hat{w_k}), \tag{3}$$

where $\otimes$ is the cross-correlation operator and R_k is the response map of the k-th features z_k.

3.2 Multi-feature fusion model

Given the target position response map generated by the tracker based on the different types of features, the goal of the integrated model is to fuse each response map R_k to obtain a stronger response map R and predict the target location. Each response map R_k can be regarded as a probability distribution with position (i, j) as the target($\sum r_k^{ij} = 1$, where r_k^{ij} represents the probability of the position (i, j) in the response map R_k, $i = \{1, 2, \ldots, N\}$, $j = \{1, 2, \ldots, M\}$). The fused response map R can reflect the consistent part of each response map R_k. The position of the maximum value in the fused response map should be considered as the predicted position of the target, so as to locate the target more accurately. Figure 2 shows the overview of the proposed multi-feature fusion model for the TIR target tracking process. To achieve this, the distribution of response map R is expected to be as close as possible to the distribution of each response map R_k. To measure the difference between the two probability distributions of each response map R_k and response map R after fusion, Jensen–Shannon (JS) divergence is adopted to measure the generalized distance between them and believes that the smaller of the distance, the smaller of the difference in their distribution. The JS divergence is a symmetric measure of the similarity of two distributions, which can give full play to the advantages of each response map [55, 56]. By minimizing the JS divergence, we can get the optimal fused response map:

$$\begin{aligned} R &= \arg\min_R \sum JS(R_k||R) \\ \text{s.t.} \quad & \sum r^{ij} = 1, \end{aligned} \tag{4}$$

where $JS(R_k||R) = \frac{1}{2}KL(R_k||M) + \frac{1}{2}KL(R||M)$, $M = \frac{1}{2}(R_k + R)$, $KL(R_k||M) = \sum r_k^{ij} \log \frac{r_k^{ij}}{m^{ij}}$, $KL(R||M) = \sum r^{ij} \log \frac{r^{ij}}{m^{ij}}$ KL denotes the Kullback–Leibler divergences, and r_k^{ij}, r^{ij}, m^{ij} represent the value of the (i, j) position in response maps R_k, R and M, respectively.

To effectively utilize the complementarity between different type of features [2, 57], we filter the response maps corresponding to a different type of features as follow:

$$R_{j,k} = R_j \odot R_k, \tag{5}$$

Equation (5) indicates that if two response maps in the same area have similar probability distributions, the filtered response map has a higher response value in that area; otherwise, it returns a lower response value. After that, Eq. (4) can be rewrite as follow:

$$\begin{aligned} R &= \arg\min_R \sum JS(R_{j,k}||R) \\ \text{s.t.} \quad & \sum r^{ij} = 1. \end{aligned} \tag{6}$$

Finding the position with the highest value in the fused response map R, and yields the tracking target location.

3.3 Model update

To adapt to the dynamic changes in the TIR target appearance during the whole tracking process, the correlation filters need to be updated continuously. We follow other correlation filters-based trackers [2, 35] who use a simple but effective linear update method to update the correlation filters:

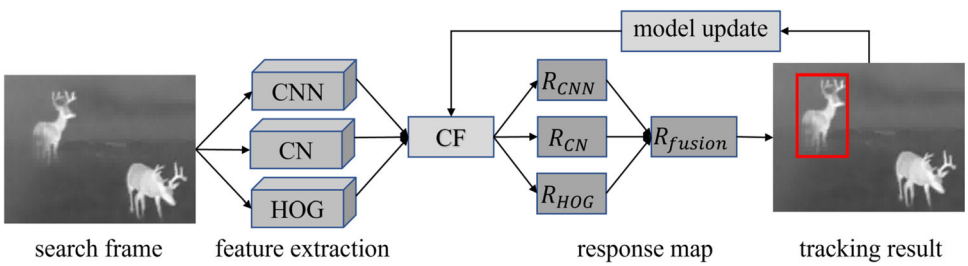

Fig. 2 The overview of the proposed multi-feature fusion model for the TIR target tracking

Table 1 Ablation studies on PTB-TIR [22] benchmark

Trackers	Hand-crafted feature	Deep feature	Precision (%)	AUC (%)	Speed (fps)
AMFT_H	✓		78.0	58.9	10.6
AMFT_D		✓	74.4	54.5	7.8
AMFT	✓	✓	81.1	61.1	7.2

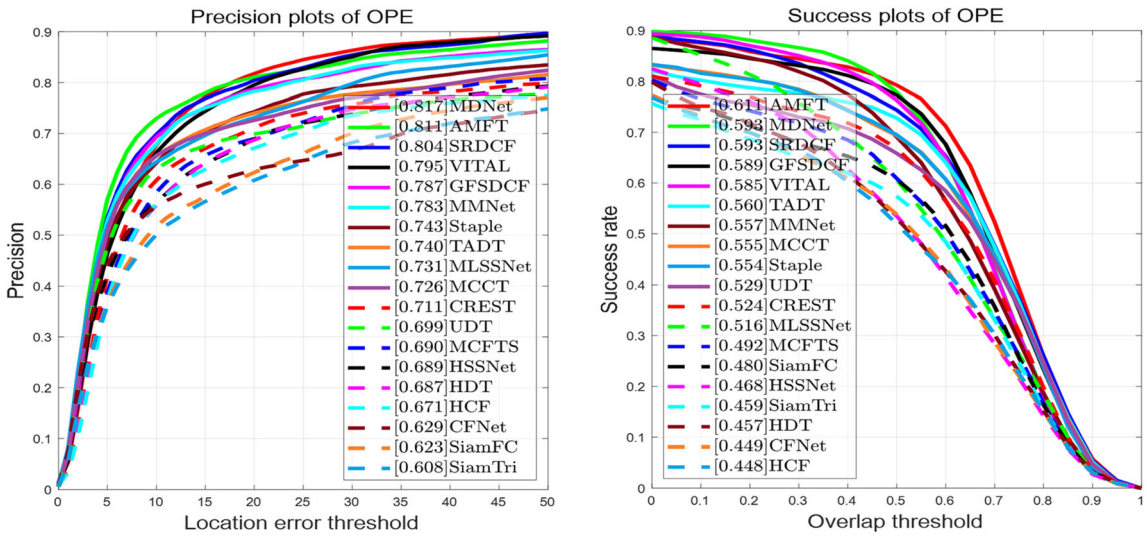

Fig. 3 Experimental comparison on PTB-TIR [22] benchmark

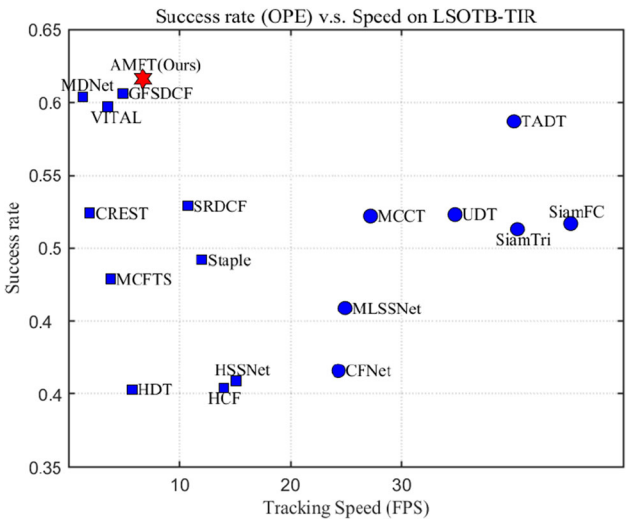

Fig. 4 Comparison results of tracking speed and tracking accuracy on the LSOTB-TIR [23] benchmark

$$w_k^t = w_k^{t-1} + \gamma w_k^t, \tag{7}$$

where γ represents the filters learning rate, w_k represents the correlation filters corresponding to the k-th features, and w_k^t represents the trained filters in the t-th frame.

4 Experiments

We verified the tracking performance of the proposed AMFT tracker on the PTB-TIR [22] and LSOTB-TIR [23] benchmarks against several other trackers, such as MCFTS [2], HSSNet [20], SianFC [43], TADT [58], MLSSNet [31], HCF [49], HDT [32], SRDCF [38], UDT [42], CFNet [44], SiamTri [45], CREST [46], VITAL [59], GFSDCF [60], MDNet [61], Staple [62], and MCCT [63]. The evaluation criteria are precision score and success score under the One Pass Evaluation (OPE) [22].

4.1 Implementation details

Experiments implemented in MATLAB2019b, and the PC is equipped with with an i7-10700-2.90GHz-CPU, and an Nvidia-GTX-1660-GPU with the matconvnet1.0-beta25 toolbox. The tracking speed of the proposed AMFT tracker is around 7 fps. The features we used in this AMFT tracker include Color Names (CN) [64], HOG [35], and deep CNN features from ResNet50 [65]. The regularization parameter $\lambda = 10\text{e}{-4}$, and the learning rate $\gamma = 10\text{e}{-2}$. The interpolation strategy has been adopted to estimate the target scale [36, 60], and it is used to predict the target location and scale with a scale factor of 7 and a scale step of 1.01.

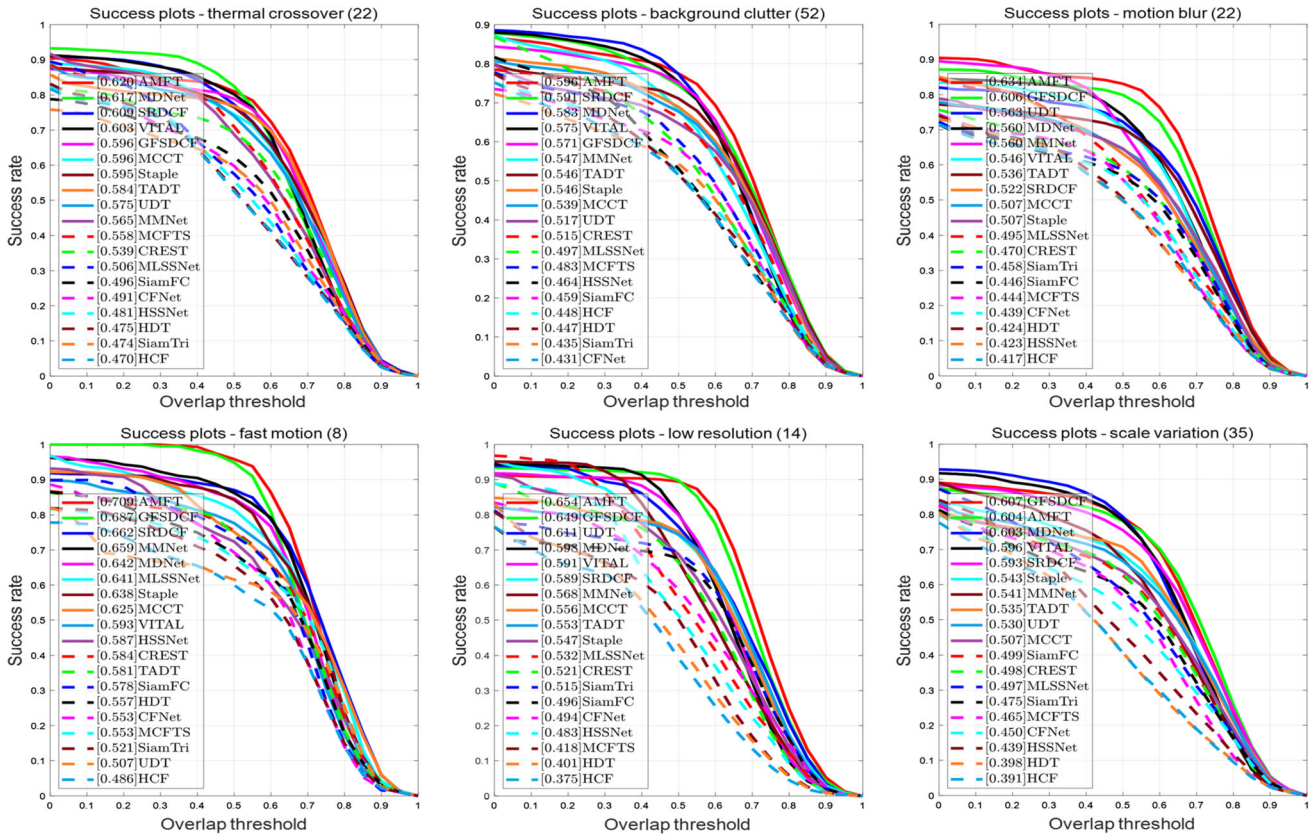

Fig. 5 Experimental comparison on PTB-TIR [22] benchmark for some attributes

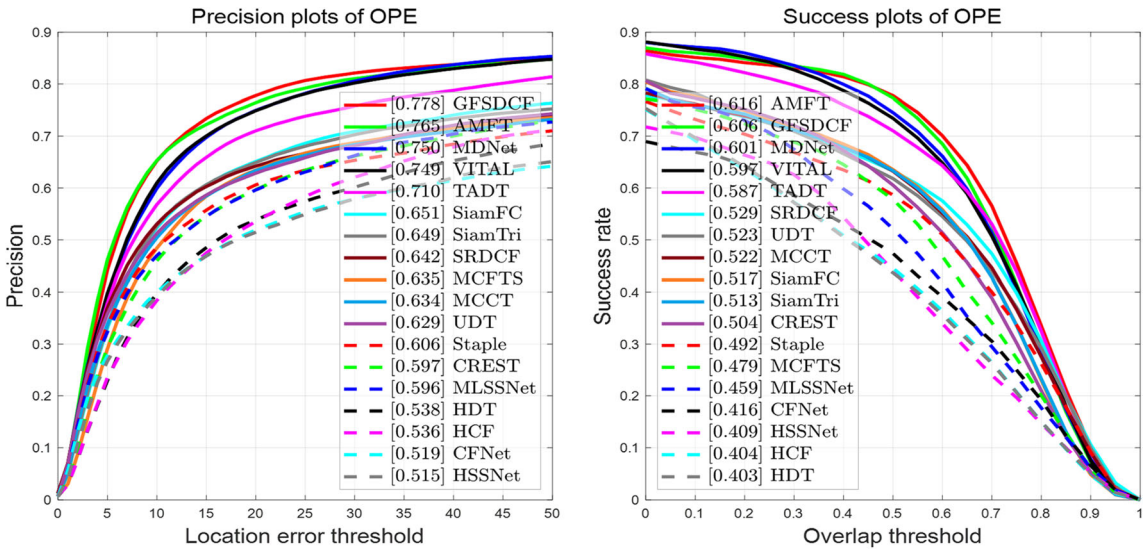

Fig. 6 Experimental comparison on LSOTB-TIR [23] benchmark

4.2 Ablation studies

In an effort to demonstrate the effectiveness of each type of the features in the proposed AMFT tracker, we provide the ablation studies using the PTB-TIR [22] benchmark. The experimental results are shown in Table 1. Note that AMFT_H represents the tracking results with only hand-crafted features, while AMFT_D represents the tracking results with only deep convolutional neural networks features. Due to the lack of color, rich texture and relatively fuzzy contour of the target in the thermal infrared image, deep features or hand-crafted features can not be used to

Table 2 Success scores (%) comparison on the LSOTB-TIR [23] benchmark for 12 different attributes, which include the scale variation (SV), fast motion (FM), motion blur (MB), distractor (DIS), low resolution (LR), intensity variation (IV), out-of-view (OV), background clutter (BC), deformation(DEF), aspect ratio variation (ARV), occlusion (OCC), and thermal crossover (TC)

Trackers	STAMT Ours	TADT [58]	MLSSNet [31]	MCCT [63]	GFSDCF [60]	SRDCF [38]	CREST [46]	Staple [62]	HSSNet [19]	HDT [32]	HCF [49]	CFNet [44]	SiamFC [43]	SiamTri [45]	MDNet [61]	UDT [42]	MCFTS [2]	VITAL [59]
SV	*69.4*	61.9	47.9	56.4	**67.1**	***62.1***	45.3	53.3	40.2	32.4	32.0	48.9	54.5	52.5	62.0	57.6	47.9	***62.1***
FM	*66.0*	***61.0***	45.8	50.2	**64.5**	51.4	51.5	44.5	40.3	44.3	41.6	39.7	57.4	56.8	57.3	53.4	47.8	58.6
MB	*63.0*	58.8	41.5	46.9	**61.6**	52.2	50.8	44.9	37.4	42.0	36.9	37.4	54.5	56.1	57.0	48.6	40.4	***59.3***
DIS	*59.4*	57.7	46.6	55.8	**58.9**	56.1	50.5	53.8	40.1	41.2	40.8	40.6	46.5	46.4	***58.3***	54.5	48.0	57.7
LR	*69.2*	65.0	51.8	53.6	**65.4**	58.1	47.7	49.4	45.8	38.9	37.1	46.2	62.3	63.7	***65.2***	57.9	36.8	60.9
IV	*70.8*	***68.6***	38.4	45.5	**69.4**	64.3	47.6	34.1	33.8	28.1	37.5	37.6	46.3	41.9	61.5	50.2	52.0	61.8
OV	*66.4*	***62.1***	47.9	54.1	**63.3**	56.9	56.5	49.3	40.7	45.0	43.0	45.9	56.0	56.6	58.2	55.4	51.3	59.7
BC	*61.0*	56.7	45.8	52.5	**59.8**	54.0	46.9	51.7	41.3	38.5	39.5	44.6	49.6	49.4	*59.4*	47.5	45.9	58.4
DEF	**57.7**	55.7	41.7	47.2	56.8	46.5	52.0	44.1	37.9	43.2	43.2	33.8	51.8	51.8	*58.0*	48.3	45.0	***57.1***
ARV	**58.5**	51.3	43.2	40.7	*59.0*	42.8	45.5	40.0	43.1	40.3	40.9	40.7	49.6	50.2	***54.9***	50.6	47.5	54.8
OCC	***57.6***	55.9	44.6	51.6	56.5	47.3	50.8	48.4	38.0	41.3	42.8	38.6	48.9	48.4	*60.8*	51.4	48.4	**59.6**
TC	51.5	51.7	32.1	***51.8***	49.5	42.5	48.4	48.5	35.0	41.7	46.0	34.8	47.7	43.4	*61.2*	41.7	43.5	**58.1**

Table 3 Success scores (%) comparison on the LSOTB-TIR [23] benchmark for 4 different scenarios, which include the handheld camera (HH), vehicle-mounted camera (VM), drone-mounted camera (DM), and surveillance camera(VS)

Trackers	AMFT Ours	TADT [58]	MLSSNet [31]	MCCT [63]	GFSDCF [60]	SRDCF [38]	CREST [46]	Staple [62]	HSSNet [20]	HDT [32]	HCF [49]	CFNet [44]	SiamFC [43]	SiamTri [45]	MDNet [61]	UDT [42]	MCFTS [2]	VITAL [59]
HH	*59.6*	***57.5***	41.1	44.8	56.5	48.0	51.9	42.4	39.1	42.2	40.8	33.9	54.0	55.3	56.8	42.2	44.0	**58.8**
VM	*76.5*	68.6	54.2	67.4	**74.6**	***74.1***	45.2	66.3	44.8	35.8	31.4	58.7	56.2	52.0	71.7	**71.1**	57.9	72.1
DM	**58.5**	53.4	44.4	44.9	*59.9*	43.8	47.3	42.0	36.9	39.6	43.3	38.4	53.0	51.6	***56.0***	43.1	45.9	54.5
VS	***57.7***	**58.0**	47.1	55.5	57.6	52.2	53.6	51.0	42.9	41.4	42.8	41.7	46.6	47.2	*59.8*	52.7	47.7	57.5

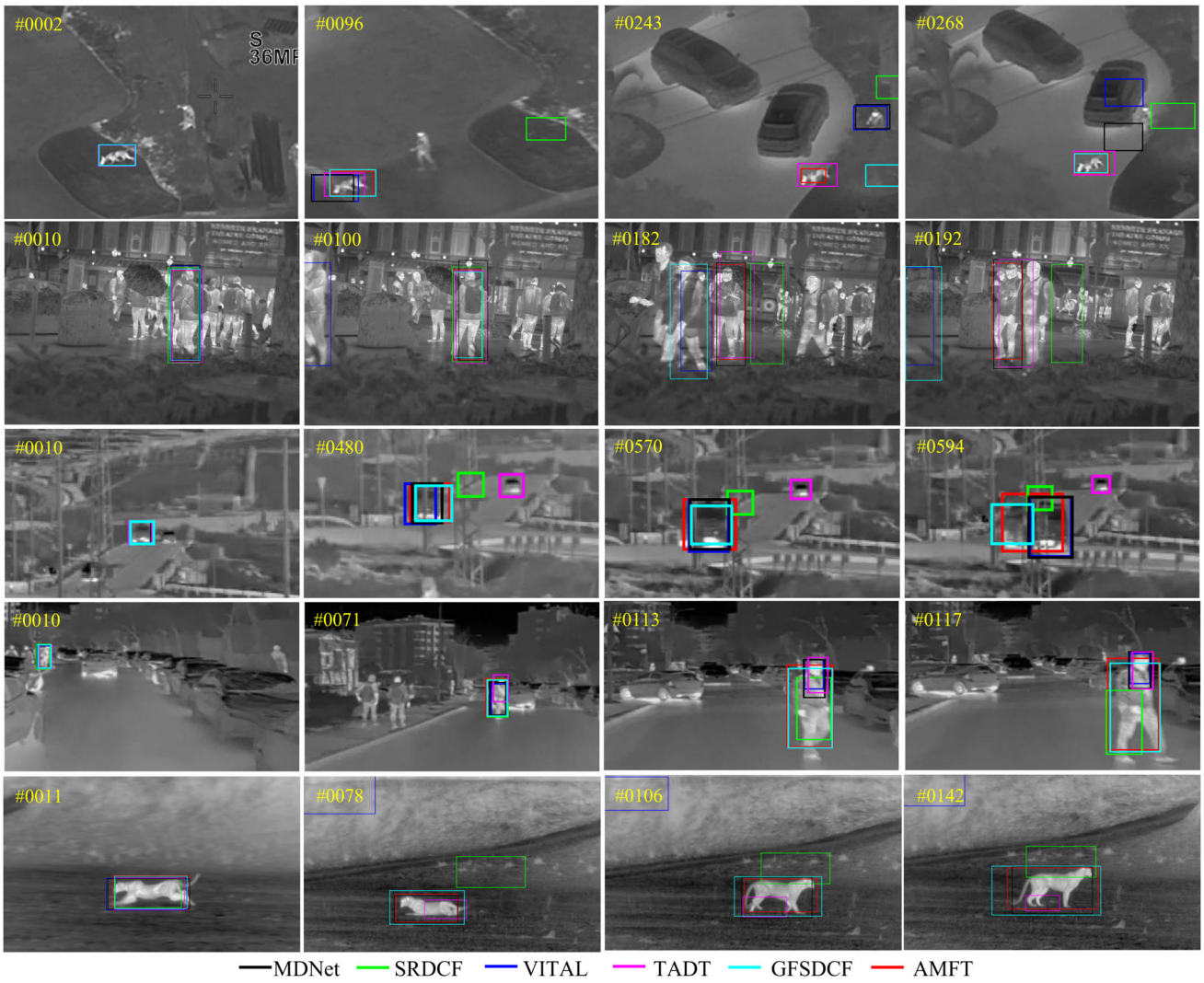

Fig. 7 Qualitative comparison of our AMFT tracking method and VITAL [59], GFSDCF [60], MDNet [61], TADT [58], SRDC4F [38] tracking methods on some TIR target tracking test video sequences(from top to bottom are dog-D-002, street-S-001, bus-S-004, person-V-007, and leopard-H-001)

represent the target well, resulting in low tracking accuracy. Table 1 shows that our AMFT tracker has significantly improved tracking performance when compared to a single-type features-based tracker. We also give the tracking speed of different types of trackers on the PTB-TIR benchmark. It can be seen from the tracking speed that multi-feature fusion will slightly increase the computation amount and reduce the tracking speed.

4.3 Comparative experiments on PTB-TIR benchmark

The experimental comparison outputs of the proposed tracker and other state-of-the-art trackers are shown in Fig. 3. We may conclude from this figure that our AMFT tracker outperformed the competition results in terms of precision and success metrics. When compared to these single-type features-based trackers [38, 43, 46, 62], our AMFT tracker performs dramatically better in terms of tracking evaluation metrics. Besides, compared with those multi-layers fused trackers [2, 20, 31, 32, 49], our tracker also achieved competitive tracking performance. Though our AMFT tracker performs somewhat worse in the precision metric than that of the MDNet [61] tracker, our AMFT tracker has a dramatically higher success score than that of the MDNet tracker, demonstrating that our tracker is more competitive than the MDNet tracker. What's more, as shown in Fig. 4 that our tracker is much faster than the MDNet tracker. The experimental results show that the multi-feature fusion model can adaptively use the complementarity between different types of features to

Fig. 8 Failure cases (from top to bottom are campus2, stranger3, and saturated). The proposed AMFT tracking results shows in red boxes and the target ground truth shows in green boxes

characterize the target appearance, which is particularly useful in the TIR target tracking task.

Figure 5 compares the performance of the proposed AMFT tracking method with that of some state-of-the-art tracking methods on the PTB-TIR [22] benchmark on some different attributes. The proposed multi-feature fusion model could further be verified to be effective in the TIR target tracking task. In comparison with these state-of-the-art trackers, our AMFT tracker has obtained good tracking results under these attributes, as shown in Fig. 5. The comparison of the thermal crossover attribute shows that our tracker could reduce the interference by other analogs. Although the success score of our AMFT tracker is lower than the GFSDCF [60] tracker on the scale variation attribute, the success score of our AMFT tracker is higher than the GFSDCF [60] tracker on the rest of the other attributes, which shows that our tracker has better tracking performance. In general, these experimental results displayed the effectiveness of our multi-feature fusion method for the TIR tracking task.

4.4 Comparative experiments on LSOTB-TIR benchmark

Figure 6 shows the tracking results comparison of our AMFT tracking method and some state-of-the-art tracking methods on the LSOTB-TIR [23] benchmark. According to Fig. 6, we know that our AMFT tracking method achieved the best success scores and the second-best precision scores. Compared with the group feature selection-based GFSDCF [60] tracker, the proposed AMFT tracker is slightly lower in the tracking precision score, but higher in tracking success precision score, which indicates that the proposed AMFT tracker achieves better performance on the LSOTB-TIR benchmark. Compared with these multi-layer deep features-based trackers (such as MFCTS [2], HDT [32], and HCF [49]), our tracker adopts the adaptive fusion strategy of hand-crafted features and deep features, which could get more accurate TIR target tracking results. Compared with the Siamese network-based trackers (such as SiamFC [43], and SiamTri [45]), our tracker obtained more than 10% improvement in each evaluation metric. Compared to other tracking methods, our AMFT tracking method also achieved better TIR target tracking results.

We compare the tracking performance of our proposed AMFT tracking method against other state-of-the-art tracking methods on some attributes and scenarios on the LSOTB-TIR [23] benchmark in order to show the tracking effectiveness of our AMFT tracking method. Table 2 shows the proposed AMFT tracking method obtained best success scores on most of the attributes (e.g., fast motion (FM), scale variation (SV), motion blur (MB), etc). For the deformation (DEF) and occlusion (OCC) attributes, the tracking success score of the proposed AMFT tracking method is lower than the MDNet [61] tracking method, probably due to the MDNet tracking method applies effective and efficient hard negative mining technology. The success score of the MDNet tracker is lower than that of our tracker in most attributes, which illustrates the effectiveness of the multiple-types of features fusion model in our AMFT tracker. In all of these tracking scenarios, our proposed AMFT tracking method received the top three success scores, as shown in Table 3. In conclusion, our proposed AMFT tracking method outperformed these state-

of-the-art tracking methods in terms of the TIR target tracking scenarios.

4.5 Qualitative comparison

Figure 7 shows the results of visual comparison between our AMFT tracking method and other state-of-the-art tracking methods on some TIR target tracking test video sequences. The MDNet [61] tracking method easily disturbed by the fast motion and scale variation attributes (e.g., dog-D-002, and person-V-007). The GFSDCF [60] tracking method gets some accurate tracking results on the dog-D-002 and leopard-H-001 test video sequences due to the group feature selection model that has been usefully adopted. However, the tracking results of the GFSDCF [60] tracking method on other test video sequences (such as street-S-001, and bus-S-004) are still unacceptable. Compared to other tracking methods, the proposed AMFT tracking method could accurately be tracking these targets in the complex tracking scenarios, which verified the proposed multi-featured fusion model is fully effective.

4.6 Failure cases

Figure 8 shows some failure cases of the proposed AMFT tracker. To display the tracking result more intuitively, we also give the ground-truth label of the target as a reference. For the stranger3 testing sequence, the main reason why the proposed AMFT tracker cannot track the target is the challenge of low resolution. For the campus2 and saturated testing sequences, due to the influence of similar distracts, our AMFT tracker shifted to other similar targets, leading to the failure of the tracking task. For these failure tracking cases, we will further explore them in future work.

5 Conclusions

In this paper, we propose a multiple types of features fusion model for the TIR target tracking task. The multi-feature fusion model adaptively integrates the hand-crafted features and the deep features by the JS divergence and gives play to their complementarity, to better model the target appearance. Meanwhile, we adopt a model update strategy to adapt to the changes of target appearance during the tracking process. Furthermore, we verify the validity of the multi-feature fusion model through the ablation studies. We demonstrate in extensive experiments on the PTB-TIR and LSOTB-TIR benchmarks that the proposed AMFT tracker has competitive tracking performance when compared to other state-of-the-art trackers.

Acknowledgements This study was supported by the National Natural Science Foundation of China (Grant Nos. 62202362, 61672183, 62172126), by the China Postdoctoral Science Foundation (Grant No. 2022TQ0247), by the Natural Science Foundation of Chongqing (Grant No.ncamc2022-msxm03), Science Foundation of The Chongqing Education Commission (Grant No.KJZD-K202200501), Foundation Project of Chongqing Normal University (Grant No.21XLB024) by the Special Research project on COVID-19 Prevention and Control of Guangdong Province (Grant No. 2020KZDZDX1227), by the Shenzhen Research Council (Grant No. JCYJ20210324120202006), by the Fundamental Research Funds for the Central Universities (Grant No. XJS222503), and by the Foundation Project of Guangzhou Institute of Technology, Xidian University (Grant No. 01131002).

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

Declarations

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

References

1. He Y-J, Li M, Zhang J, Yao J-P (2015) Infrared target tracking via weighted correlation filter. Infrared Phys Technol 73:103–114
2. Liu Q, Lu X, He Z, Zhang C, Chen W-S (2017) Deep convolutional neural networks for thermal infrared object tracking. Knowl Based Syst 134:189–198
3. Wang Y, Wei X, Tang X, Wu J, Fang J (2022) Response map evaluation for RGBT tracking. Neural Comput Appl 34(7):5757–5769
4. Gundogdu E, Koc A, Solmaz B, Hammoud RI, Aydin Alatan A (2016) Evaluation of feature channels for correlation-filter-based visual object tracking in infrared spectrum. In: CVPRW, IEEE, pp 24–32
5. Lamberti F, Sanna A, Paravati G (2011) Improving robustness of infrared target tracking algorithms based on template matching. IEEE Trans Aerosp Electron Syst 47(2):1467–1480
6. Chen J, Lin Y, Huang D, Zhang J (2020) Robust tracking algorithm for infrared target via correlation filter and particle filter. Infrared Phys Technol 111:103516103516
7. He Y, Li M, Zhang J, Yao J (2015) Infrared target tracking based on robust low-rank sparse learning. IEEE Geosci Remote Sens Lett 13(2):232–236
8. Yuan D, Chang X, Liu Q, Wang D, He Z (2021) Active learning for deep visual tracking. arXiv preprint arXiv:2110.13259
9. Wang P, Sun M, Wang H, Li X, Yang Y (2020) Convolution operators for visual tracking based on spatial-temporal regularization. Neural Comput Appl 32(10):5339–5351
10. Song X, Jin Z (2022) Robust label rectifying with consistent contrastive-learning for domain adaptive person re-identification. IEEE Trans Multimedia 24:3229–3239
11. Shu X, Yang Y, Wu B (2021) A neighbor level set framework minimized with the split Bregman method for medical image segmentation. Signal Process 189:108293
12. Li R, Zhang B, Kang D-J, Teng Z (2019) Deep attention network for person re-identification with multi-loss. Comput Electr Eng 79:106455

13. Yuan D, Fan N, He Z (2020) Learning target-focusing convolutional regression model for visual object tracking. Knowl Based Syst 194:105526
14. Shu X, Yang Y, Wu B (2021) Adaptive segmentation model for liver CT images based on neural network and level set method. Neurocomputing 453:438–452
15. Song X, Jin Z (2022) Domain adaptive attention-based dropout for one-shot person re-identification. Int J Mach Learn Cybern 13(1):255–268
16. Yan C, Chang X, Li Z, Guan W, Ge Z, Zhu L, Zheng Q (2021) Zeronas: differentiable generative adversarial networks search for zero-shot learning. IEEE Trans Pattern Anal Mach Intell 41:1–9
17. Gao P, Ma Y, Song K, Li C, Wang F, Xiao L (2018) Large margin structured convolution operator for thermal infrared object tracking. In: ICPR, IEEE, pp 2380–2385
18. Liu Q, Li X, He Z, Fan N, Yuan D, Liu W, Liang Y (2020) Multi-task driven feature models for thermal infrared tracking. In: AAAI, vol 34, AAAI, pp 11604–11611
19. Zhang L, Gonzalez-Garcia A, Van De Weijer J, Danelljan M, Khan FS (2018) Synthetic data generation for end-to-end thermal infrared tracking. IEEE Trans Image Process 28(4):1837–1850
20. Li X, Liu Q, Fan N, He Z, Wang H (2019) Hierarchical spatial-aware Siamese network for thermal infrared object tracking. Knowl Based Syst 166:71–81
21. Li M, Peng L, Chen Y, Huang S, Qin F, Peng Z (2019) Mask sparse representation based on semantic features for thermal infrared target tracking. Remote Sens 11(17):1967
22. Liu Q, He Z, Li X, Zheng Y (2019) PTB-TIR: a thermal infrared pedestrian tracking benchmark. IEEE Trans Multimedia 22(3):666–675
23. Liu Q, Li X, He Z, et al (2020) LSOTB-TIR: a large-scale high-diversity thermal infrared object tracking benchmark. In: ACM MM, ACM, pp 3847–3856
24. Li R, Zhang B, Teng Z, Fan J (2022) An end-to-end identity association network based on geometry refinement for multi-object tracking. Pattern Recogn 129:108738
25. Marvasti-Zadeh SM, Ghanei-Yakhdan H, Kasaei S (2021) Efficient scale estimation methods using lightweight deep convolutional neural networks for visual tracking. Neural Comput Appl 33(14):8319–8334
26. Yuan D, Kang W, He Z (2020) Robust visual tracking with correlation filters and metric learning. Knowl Based Syst 195:105697
27. Dawoud A, Alam MS, Bal A, Loo C (2006) Target tracking in infrared imagery using weighted composite reference function-based decision fusion. IEEE Trans Image Process 15(2):404–410
28. Yuan D, Chang X, Li Z, He Z (2021) Learning adaptive spatial-temporal context-aware correlation filters for UAV tracking. ACM Trans Multimed Comput Commun Appl 18(3):70:1-70:18
29. Yu T, Mo B, Liu F, Qi H, Liu Y (2019) Robust thermal infrared object tracking with continuous correlation filters and adaptive feature fusion. Infrared Phys Technol 98:69–81
30. Li G, Peng M, Nai K, Li Z, Li K (2020) Multi-view correlation tracking with adaptive memory-improved update model. Neural Comput Appl 32(13):9047–9063
31. Liu Q, Li X, He Z, Fan N, Yuan D, Wang H (2021) Learning deep multi-level similarity for thermal infrared object tracking. IEEE Trans Multimedia 23:2114–2126
32. Qi Y, Zhang S, Qin L, Yao H, Huang Q, Lim J, Yang M-H (2016) Hedged deep tracking. In: CVPR, IEEE, pp 4303–4311
33. Yuan D, Zhang X, Liu J, Li D (2019) A multiple feature fused model for visual object tracking via correlation filters. Multimedia Tools Appl 78(19):27271–27290
34. Li X, Huang L, Wei Z, Nie J, Chen Z (2021) Adaptive multi-branch correlation filters for robust visual tracking. Neural Comput Appl 33(7):2889–2904
35. Henriques JF, Caseiro R, Martins P, Batista J (2014) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596
36. Kiani Galoogahi H, Fagg A, Lucey S (2017) Learning background-aware correlation filters for visual tracking. In: ICCV, IEEE, pp 1135–1143
37. Yuan D, Shu X, He Z (2020) TRBACF: learning temporal regularized correlation filters for high performance online visual object tracking. J Vis Commun Image Rep 72:102882
38. Danelljan M, Hager G, Shahbaz Khan F, Felsberg M (2015) Learning spatially regularized correlation filters for visual tracking. In: ICCV, IEEE, pp 4310–4318
39. Bibi A, Mueller M, Ghanem B (2016) Target response adaptation for correlation filter tracking. In: ECCV, Springer, pp 419–433
40. Yuan D, Li X, He Z, Liu Q, Lu S (2020) Visual object tracking with adaptive structural convolutional network. Knowl Based Syst 194:105554
41. Yang K, Song H, Zhang K, Liu Q (2020) Hierarchical attentive Siamese network for real-time visual tracking. Neural Comput Appl 32(18):14335–14346
42. Wang N, Song Y, Ma C, Zhou W, Liu W, Li H (2019) Unsupervised deep tracking. In: CVPR, IEEE, pp 1308–1317
43. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH (2016) Fully-convolutional Siamese networks for object tracking. In: ECCV, Springer, pp 850–865
44. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr PH (2017) End-to-end representation learning for correlation filter based tracking. In: CVPR, IEEE, pp 2805–2813
45. Dong X, Shen J (2018) Triplet loss in Siamese network for object tracking. In: ECCV, Springer, pp 459–474
46. Song Y, Ma C, Gong L, Zhang J, Lau RW, Yang M-H (2017) CREST: convolutional residual learning for visual tracking. In: ICCV, IEEE, pp 2574–2583
47. Li R, Zhang B, Teng Z, Fan J (2021) A divide-and-unite deep network for person re-identification. Appl Intell 51(3):1479–1491
48. Yuan D, Shu X, Liu Q, He Z (2022) Structural target-aware model for thermal infrared tracking. Neurocomputing 491:44–56
49. Ma C, Huang J-B, Yang X, Yang M-H (2015) Hierarchical convolutional features for visual tracking. In: ICCV, IEEE, pp 3074–3082
50. Li M, Cai W, Verspoor K, Pan S, Liang X, Chang X (2022) Cross-modal clinical graph transformer for ophthalmic report generation. In: CVPR, pp 20656–20665
51. Elayaperumal D, Joo YH (2021) Robust visual object tracking using context-based spatial variation via multi-feature fusion. Inf Sci 577:467–482
52. Li M, Huang P-Y, Chang X, Hu J, Yang Y, Hauptmann A (2022) Video pivoting unsupervised multi-modal machine translation. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2022.3181116
53. Zhang L, Danelljan M, Onzalez-Garcia A, van de Weijer J, Shahbaz Khan F (2019) Multi-modal fusion for end-to-end rgb-t tracking. In: ICCVW, IEEE, pp 2252–2261
54. Li C, Lu A, Zheng A, Tu Z, Tang J (2019) Multi-adapter RGBT tracking. In: ICCVW, IEEE, pp 2262–2270
55. Wang F, Vemuri BC, Rangarajan A (2006) Groupwise point pattern registration using a novel CDF-based Jensen–Shannon divergence. In: CVPR, IEEE, pp 1283–288
56. Sutter T, Daunhawer I, Vogt JE (2020) Multimodal generative learning utilizing Jensen–Shannon divergence. In: NeurIPS, Curran, pp 6100–6110
57. Li X, Liu Q, He Z, Wang H, Zhang C, Chen W-S (2016) A multi-view model for visual tracking via correlation filters. Knowl Based Syst 113:88–99
58. Li X, Ma C, Wu B, He Z, Yang M-H (2019) Target-aware deep tracking. In: CVPR, IEEE, pp 1369–1378

59. Song Y, Ma C, Wu X, Gong L, Bao L, Zuo W, Shen C, Lau RW, Yang M-H (2018) Vital: visual tracking via adversarial learning. In: CVPR, IEEE, pp 8990–8999
60. Xu T, Feng Z-H, Wu X-J, Kittler J (2019) Joint group feature selection and discriminative filter learning for robust visual object tracking. In: ICCV, IEEE, pp 7950–7960
61. Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: CVPR, IEEE, pp 4293–4302
62. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr PH (2016) Staple: complementary learners for real-time tracking. In: CVPR, IEEE, pp 1401–1409
63. Wang N, Zhou W, Tian Q, Hong R, Wang M, Li H (2018) Multi-cue correlation filters for robust visual tracking. In: CVPR, IEEE, pp 4844–4853
64. Danelljan M, Shahbaz Khan F, Felsberg M, Van de Weijer J (2014) Adaptive color attributes for real-time visual tracking. In: CVPR, IEEE, pp 1090–1097
65. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, IEEE, pp 770–778