remote sensing

MDPI

Article

SiamCAF: Complementary Attention Fusion-Based Siamese Network for RGBT Tracking

Yingjian Xue [1], Jianwei Zhang [1,*], Zhoujin Lin [1], Chenglong Li [1], Bihan Huo [1] and Yan Zhang [2]

[1] School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211215007@nuist.edu.cn (Y.X.); 20211215025@nuist.edu.cn (Z.L.); 202212150015@nuist.edu.cn (C.L.); 20211215022@nuist.edu.cn (B.H.)

[2] Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211220042@nuist.edu.cn

* Correspondence: zhangjw@nuist.edu.cn

Abstract: The tracking community is increasingly focused on RGBT tracking, which leverages the complementary strengths of corresponding visible light and thermal infrared images. The most well-known RGBT trackers, however, are unable to balance performance and speed at the same time for UAV tracking. In this paper, an innovative RGBT Siamese tracker named SiamCAF is proposed, which utilizes multi-modal features with a beyond-real-time running speed. Specifically, we used a dual-modal Siamese subnetwork to extract features. In addition, to extract similar features and reduce the modality differences for fusing features efficiently, we designed the Complementary Coupling Feature fusion module (CCF). Simultaneously, the Residual Channel Attention Enhanced module (RCAE) was designed to enhance the extracted features and representational power. Furthermore, the Maximum Fusion Prediction module (MFP) was constructed to boost performance in the response map fusion stage. Finally, comprehensive experiments on three real RGBT tracking datasets and one visible–thermal UAV tracking dataset showed that SiamCAF outperforms other tracking methods, with a remarkable tracking speed of over 105 frames per second.

Keywords: multi-modal object tracking; RGBT tracking; attention mechanism; deep learning

Citation: Xue, Y.; Zhang, J.; Lin, Z.; Li, C.; Huo, B.; Zhang, Y. SiamCAF: Complementary Attention Fusion-Based Siamese Network for RGBT Tracking. *Remote Sens.* **2023**, *15*, 3252. https://doi.org/10.3390/rs15133252

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 27 April 2023
Revised: 8 June 2023
Accepted: 22 June 2023
Published: 24 June 2023

1. Introduction

Visual object tracking is not only a consequential but also fundamental task in the realm of computer vision, requiring the accurate and robust tracking of objects across subsequent frames based on the initial position of a model-agnostic target. This technology is helpful for potential practical applications such as visual monitoring, robot vision navigation, and autonomous vehicles. The performance of object tracking tends to degrade under challenging circumstances, such as low illumination, rainy, fog, and other extreme environments, due to the inherent limitations of visible light images. As shown in Figure 1, under certain conditions such as low light and partial occlusion, the targets may lack clear distinguishability in visible light images. On the other hand, thermal infrared images may offer a more distinct and discernible representation of the targets. Conversely, in scenarios such as inadequate thermal imaging of the target or thermal interferences, the visual information, including color and texture, that is captured in visible light images can effectively display the target, as demonstrated in Figure 2.

The thermal image obtained with the thermal infrared camera, as a complementary cue, can effectively compensate for the degradation in object-tracking performance, and thermal infrared cameras have become increasingly affordable and economically accessible in recent years [1]. An increasing number of RGBT tracking benchmark datasets [2–5] serve as a versatile evaluation platform for assessing the performance of trackers. This has contributed to increasing attention and interest in RGBT tracking as a research area in computer vision.

Figure 1. Examples where objects are tracked better in thermal infrared images (**bottom**).

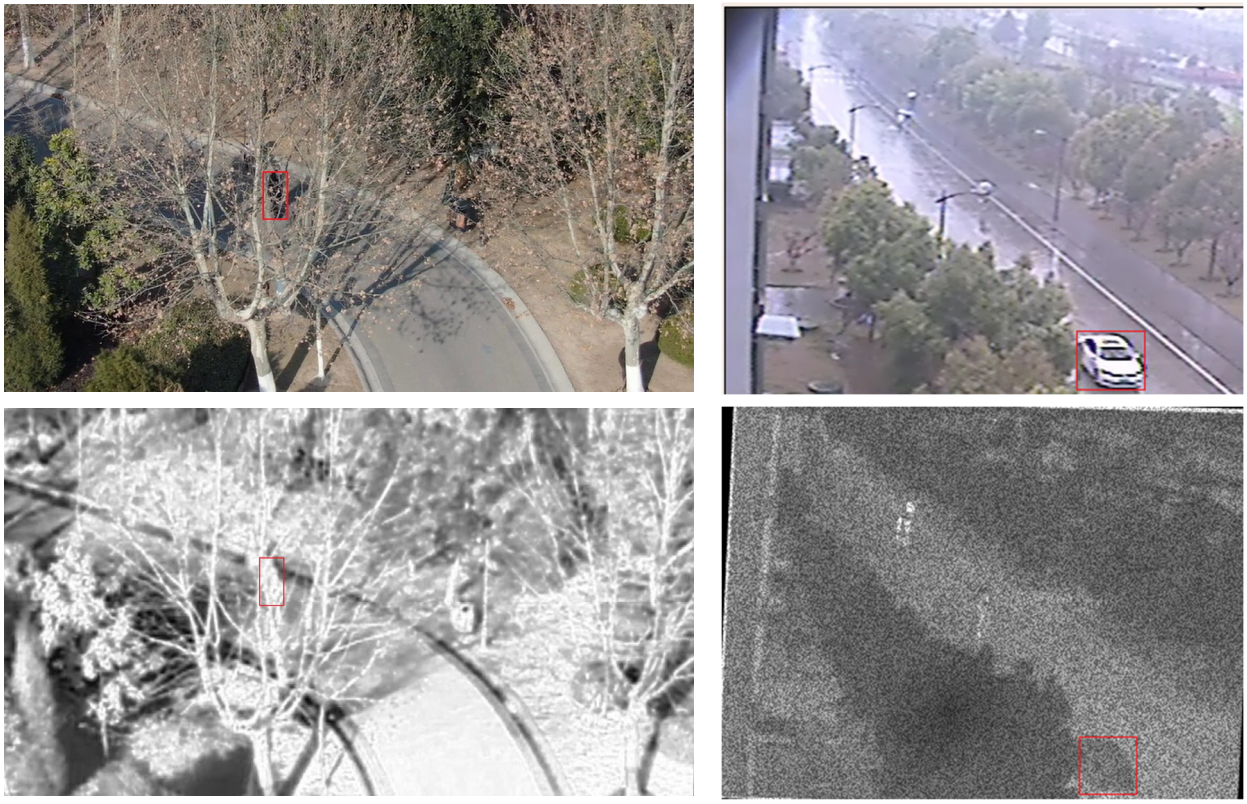


Figure 2. Examples where objects are tracked better in visible light images (**top**).

Thus far, a great number of RGBT trackers have been proposed. Traditional object-tracking methods in their early stages of development, such as Kalman filter [6], particle filter [7], and mean shift [8], were used for RGBT tracking. These methods mostly use hand-crafted features such as a histogram of oriented gradients (HOG) [9], scale invariant feature transform (SIFT) [10], and local binary pattern (LBP) [11] to fuse features. Nevertheless, traditional RGBT tracking methods do have certain limitations that can restrict their overall performance. First of all, the features used in tracking are handcrafted, meaning that they cannot handle practical challenges such as scale changes and fast movements. Secondly, the above methods often necessitate significant computational resources, and almost none of them are capable of meeting demanding real-time requirements. Influenced by the notable achievements of Convolutional Neural Networks in visible light tracking, an increasing number of attempts have been made to use CNNs in order to enhance the performance of RGBT trackers. The Multi-Domain Network (MDNet) [12] and Siamese [13] architectures are two widely used and popular frameworks in the realm of RGBT tracking. MDNet-based trackers must be disregarded due to their slower processing speeds that do not meet real-time requirements, while trackers based on the Siamese network satisfy the real-time requirements. Despite the faster speed achieved with Siamese-based trackers, there still remains a large performance gap when compared to most advanced RGBT trackers.

These approaches can effectively take advantage of the modality characteristic, but the majority of them are missing the potential benefits of modality differences between visible light and thermal infrared features, which are vital for the adequate fusion of different modalities. At the same time, the question of how to strike a balance between the high performance and high speed of RGBT trackers is also a meaningful and challenging issue that necessitates further exploration.

In this paper, we propose a Siamese Complementary Attention Fusion network (SiamCAF) which can achieve a high performance and above-real-time speed. We first expanded the Siamese framework to a dual Siamese framework [14] to extract different features from corresponding images of two modalities. Afterwards, the extracted features were fused through the Complementary Coupling Feature fusion module (CCF), which extracts the similar features through coupled filters to reduce the modality differences and then enhances the discriminative power of the fused features. The visible light and thermal infrared features utilize the Residual Channel Attention Enhanced module (RCAE) to achieve the feature enhancement of the respective modalities. Finally, the Maximum Fusion Prediction module (MFP), employed for fusing three predicted position maps, was utilized to accomplish the final fusion at the response level.

The primary contributions of this research paper can be summarized as follows:

1. We extended the Siamese network to RGBT tracking for better utilization of the information of two modalities. As a result, our proposed method demonstrates an outstanding performance and speed (105 FPS), surpassing most advanced RGBT trackers based on the current mainstream datasets.
2. We designed a Complementary Coupling Feature fusion module (CCF) which can extract similar features and reduce modality differences to fuse features better. Simultaneously, the features are enhanced using the Residual Channel Attention Enhanced module (RCAE) to amplify the characteristics of visible light and thermal infrared modalities.
3. We propose a Maximum Fusion Prediction module (MFP) in the response map fusion stage which enables us to effectively accomplish the response level fusion.

The subsequent sections of this paper are structured as follows: Section 2 provides a comprehensive review of the relevant domestic and international studies related to our approach. Section 3 presents the details of our tracking network. Section 4 describes our implementation details and describes the experimental results in mainstream datasets such as GTOT, RGBT234, and VTUAV. Section 5 draws the conclusion.

2. Related Works

2.1. Visual Object Tracking

Modern visual object tracking can roughly be categorized into two main branches. The first branch of tracking approaches are predicated on the correlation filter, which involves training a regressor by diagonalizing the resulting data matrix with Discrete Fourier Transform. This approach enables concurrent online tracking and weight updates of the filters. The concept of using correlation filters for object tracking was initially introduced with MOSSE [15], and it has since been widely used for tracking. Henriques et al. [16] proposed CSK to address the issue of insufficient samples in MOSSE. KCF [17] augments the previous single-channel correlation filter by defining the multi-channel connection mode. Subsequent correlation filter methods use more features, such as color features [18] and deep features [19], to enhance tracking accuracy, but these additional features often result in a significant reduction in the model update speed. The tracking approaches in the other branch are built upon deep learning. As an example, MDNet [12] is the pioneering masterpiece among the early tracking algorithms based on CNNs whose core idea is to use network branches with multi-domains to fit different target objects. SiamRPN [20] constructs an RPN structure based on the Siamese network [13]. The template frame and the detection area use the same network to extract features and determine the location and size of the target through two independent network branches: classification and regression.

2.2. RGBT Object Tracking

As the theoretical research on thermal infrared cameras improves and a growing number of tracking benchmarks are proposed [2–5], the field of RGBT object tracking is recently attracting a great deal of attention.

A crucial concern in RGBT tracking is how to optimally exploit the information from both modalities, allowing them to synergistically complement each other for an improved tracking performance. Zhang et al. [21] introduced the fusion-based approach that combines visible light and thermal infrared images, followed by tracking based on the fused image data. In SiamFT [22], convolutional features extracted from images of two modalities are concatenated to generate fused features. The cross-relation operation is then employed on them to generate the ultimate response map. DSiamMFT [23] incorporates the dynamic online learned transformation strategy and multi-level semantic features, building upon the method established in SiamFT. In DuSiamRT [14], a response-level fusion-tracking algorithm is proposed that incorporates deep learning techniques. Additionally, it incorporates the weight distribution mechanism during the feature extraction stage, further enhancing the tracking performance. SiamCDA [24] fuses the cross-modal information based on SiamRPN++ and takes the influences of distractors into consideration. Feng et al. [25] proposed a pioneering framework based on Transformer, designing a simple Siamese network to extract features which are then input into the Transformer feature fusion network to complete target tracking. SiamIVFN [26] is a fusion tracker whose tracking head is built based on SiamFC++. Real-time tracking is always considered in the design of SiamIVFN models; thus, the structure of SiamIVFN is straightforward.

2.3. Attention Mechanisms

Attention mechanisms, first introduced for machine translation, have now become an essential concept within the realm of whole computer vision. They have gained significant attention and have become a crucial concept in recent research efforts. Many works have explored the significance of attention mechanisms in different spatial and channel domains to improve performance in the main task. Attention mechanisms play a vital role in enabling neural networks to only concentrate on important details, avoiding unimportant information, much like our visual processing system. This selective attention aids in perception by allowing the network to prioritize important features, similar to the way in which humans tend to focus on specific parts of an image while disregarding irrelevant details. The spatial transformation of the input image was proposed with STN [27] to

strengthen the model's capacity for generalization and robustness. SENet [28] is a method that learns channel-wise correlations to allocate more attention to channels with higher information contents. This allows the network to dynamically adapt its attention to different channels based on their relevance. CBAM [29] emphasizes the meaningful features in channel and spatial dimensions and applies attention modules in turn to learn what to focus on and where. SKNet [30] is an attention mechanism research project on convolution kernels that uses different convolutional kernel weights for different images.

3. Our Method

In this section, we provide a detailed introduction to our proposed RGBT tracking model. First, we outline the overall network architecture of the SiamCAF. Then, we describe each component module in detail. As depicted in Figure 3, the network has a dual-modal Siamese subnetwork to extract features, CCF modules to fuse dual-modal features, RCAE modules for unimodal feature enhancement, region proposal networks for proposal generation, and an MFP module to select the bounding box. In the following sections, we comprehensively elucidate each component in detail.

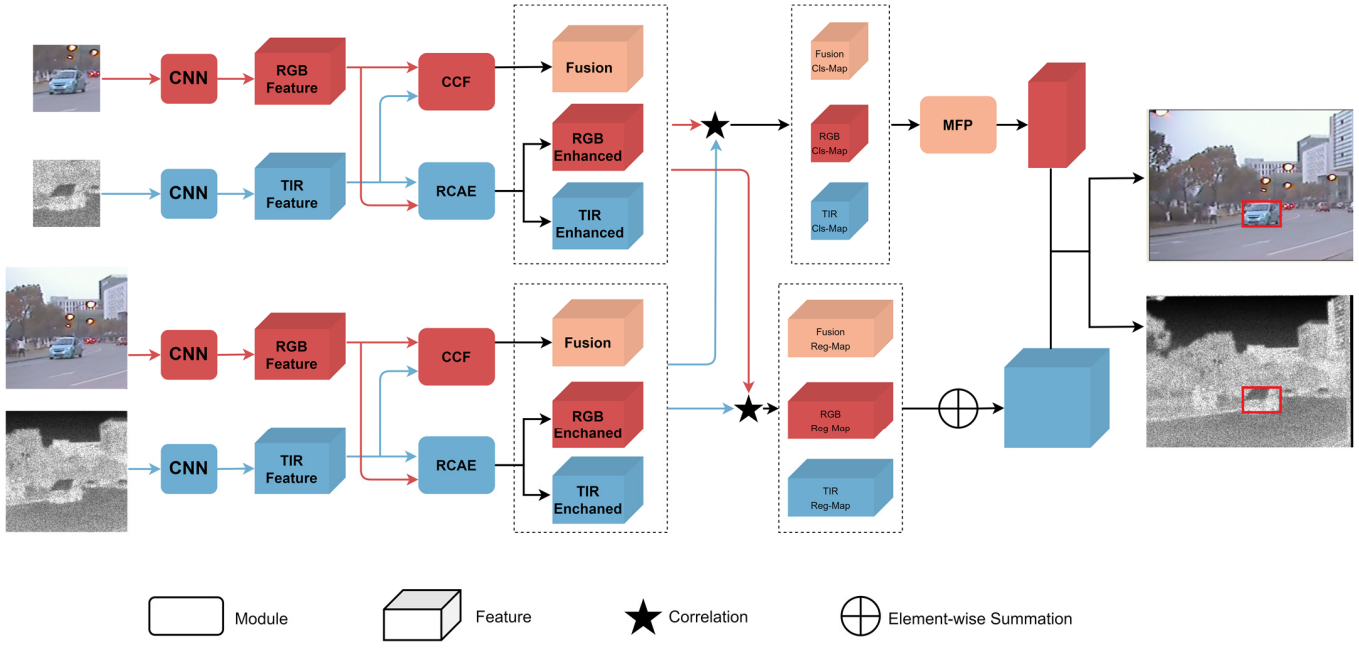


Figure 3. Pipeline of our proposed SiamCAF.

3.1. Dual-Modal Siamese Subnetwork for Feature Extraction

We propose the tracking model SiamCAF, which has two Siamese subnetworks [14], named RGB Siamese network and T Siamese network. They are utilized for distinctive feature extraction from visible light and thermal infrared images. For the better processing of features by subsequent modules, the two Siamese subnetworks use an identical structure but possess distinct parameters that enable them to accurately perform feature extraction from the corresponding image pairs. Each Siamese subnetwork contains two branches, namely, the template branch and detection branch, which have the same structure and parameters. However, the difference is that the template branch is responsible for extracting features from the template patches. We denote the template patches corresponding to the visible light and thermal infrared images as z_r and z_t. The detection branch extracts features from the detection patches. We denote the detection patches corresponding to the visible light and thermal infrared images as x_r and x_t. The template patches and detection patches are obtained by cropping regions of interest from the template frames and detection frames, respectively. The first frame of the tracked object is referred to as the template frame, and the subsequent frames which need to be tracked are called the detection frames. The

convolutional neural network utilized in our approach is an improved version of AlexNet, which removes padding in the same way as SiamRPN [20]. For ease of notation, we denote the feature extraction operations of the RGB Siamese network and T Siamese network as $\varphi_r(\cdot)$ and $\varphi_t(\cdot)$. Then, the dual-modal Siamese subnetwork's output includes $\varphi_r(z_r)$, $\varphi_r(x_r)$, $\varphi_t(z_t)$, and $\varphi_t(x_t)$.

3.2. Complementary Coupling Feature Fusion Module

Given the extracted features from the Siamese networks of two modalities, the issue of how to fuse them in a more efficient form is the next question. Similar to the existing RGBT trackers, the features $\varphi_r(z_r)$ from the RGB Siamese network in the template branch and the matching features $\varphi_t(z_t)$ extracted from the T Siamese network in the template branch are combined to generate the fused template features. Likewise, $\varphi_r(x_r)$ and $\varphi_t(x_t)$ in the detection branch are combined to gain the fused detection features. The most straightforward and commonly used methods for multimodal features fusion are element-wise summation [23] and concatenation [31]; nevertheless, they do not take the differences between features into consideration. Specifically, simple element-wise summation and concatenation do not take the characteristics and reliability of different modalities into account, which, indeed, is why the fusion strategy based on content dependency weighting often yields a superior performance. Despite this, the vast majority of the current fusion strategies lack consideration of the dissimilarities between features of the two modalities.

Based on the above analysis, as depicted in Figure 4, we propose a multimodal fusion module that integrates visible light and thermal infrared features to improve discriminability, called the Complementary Coupling Feature fusion module (CCF). Inspired by [32], we first used coupled filters with a coupling ratio of 0.5 in the convolutional layer to perform the extraction of similar features between the visible light and thermal infrared features. The upper red part represents the non-coupling filter of visible light features and the lower gray part represents the non-coupling filter of thermal infrared features. The overlapping yellow part between the two indicates the coupled part of the two filters. In this way, the weight of visible light and thermal infrared features can be updated using the coupled filters at the same time. In each iteration, the non-coupling filter is updated once, and the coupled filter is updated twice. All the non-coupling filters and coupled filters in the convolutional layer, with a kernel size of 3×3, produce two weight maps. These weight maps are then normalized to the range of $[0, 1]$ using a sigmoid layer, indicating the extent to which additional information from one modality feature needs to be incorporated into another. Taking the template branch as an example, the weight maps can be obtained as follows:

$$W_r = \sigma(conv(\varphi_r(z_r), \theta_1)) \tag{1}$$

$$W_t = \sigma(conv(\varphi_t(z_t), \theta_2)) \tag{2}$$

where $conv(*, \theta)$ denotes the convolutional layer with the parameters θ, including both non-coupling filters and coupled filters, and the parameters of the coupled filters are the same. $\sigma(\cdot)$ denotes the sigmoid layer.

After we obtain the weight maps, the visible light and thermal infrared features are enhanced through cross-modal connections with W_r and W_t, and the enhanced features $\varphi_r^{'}(z_r)$ and $\varphi_t^{'}(z_t)$ can be obtained as follows:

$$\varphi_r^{'}(z_r) = \varphi_r(z_r) + \varphi_t(z_t) \otimes W_t \tag{3}$$

$$\varphi_t^{'}(z_t) = \varphi_t(z_t) + \varphi_r(z_r) \otimes W_r \tag{4}$$

where $\otimes$ denotes the element-wise multiplication. Since the enhanced features $\varphi_r^{'}(z_r)$ and $\varphi_t^{'}(z_t)$ contain information about another modality, the difference between $\varphi_r^{'}(z_r)$ and $\varphi_t^{'}(z_t)$ is smaller than before.

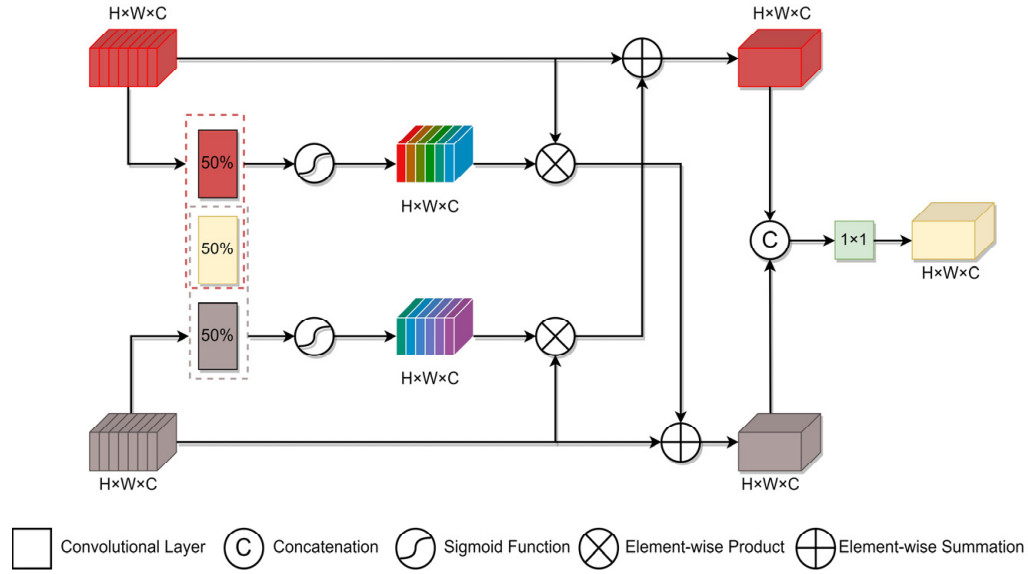


Figure 4. Illustration of CCF. CCF first extracts the similar features between visible light and thermal infrared features using the coupled filter to obtain the weight maps. Then, it enhances the features using cross-modal connections. Finally, CCF fuses the enhanced features via concatenation and uses the 1×1 convolutional layer to fuse the channel information.

Finally, we fuse the enhanced features through concatenation, and the channel information is fused using the 1×1 convolutional layer. Then, the fused features z_f can be obtained as follows:

$$z_f = conv\Big(cat\Big(\varphi_r'(z_r), \varphi_t'(z_t)\Big), \theta_3\Big) \tag{5}$$

where $cat(\cdot)$ denotes the concatenation operation and $conv(*, \theta_3)$ denotes the convolutional layer with a kernel size 1×1 and parameters θ_3.

3.3. Residual Channel Attention Enhanced Module

To fully utilize visible light and thermal infrared features while suppressing feature noise and redundancy, inspired by SENet [28], which dynamically recalibrates the feature responses of each channel, we developed a Residual Channel Attention Enhanced module (RCAE). The features $\varphi_r(z_r)$ from the RGB Siamese network in the template branch, along with the corresponding features $\varphi_t(z_t)$ from the T Siamese network in the template branch, are enhanced together via the RCAE. Similarly, $\varphi_r(x_r)$ and $\varphi_t(x_t)$ in the detection branch are enhanced in the same way as those in the template branch. The importance of each feature channel is acquired automatically through the learning process, allowing for the promotion of useful features and the suppression of features that are not relevant for the current task based on their importance scores, thus enhancing the representation capabilities of the network by fully magnifying the characteristics of different modalities.

As shown in Figure 5, RCAE concatenates the original features which are extracted from visible light and thermal infrared images for better information interaction. We use global average pooling to squeeze the global spatial information into a channel descriptor. Taking the template branch as an example, formally, a statistic $g \in R^c$ is generated by shrinking the features through its spatial dimensions $H \times W$, where the c-th element of g is computed as follows:

$$z_{r+t} = cat(\varphi_r(z_r), \varphi_t(z_t)) \tag{6}$$

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (z_{r+t})_c(i, j) \tag{7}$$

where $cat(\cdot)$ denotes the concatenation operation. The global feature then passes through two fully connected layers to improve the generalization ability of RCAE, and the subsequent sigmoid layer is used to normalize the values of the global feature to $[0, 1]$:

$$h_c = \sigma(\beta(\alpha(g_c))) \tag{8}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ denote two different fully connected layers and $\sigma(\cdot)$ denotes one sigmoid layer. The learned feature vector h_c is multiplied by the original feature z_{r+t} and then added to the original feature to calculate the output z:

$$z_c = (z_{r+t})_c \cdot h_c + (z_{r+t})_c = cat(\varphi_r^*(z_r), \varphi_t^*(z_t)) \tag{9}$$

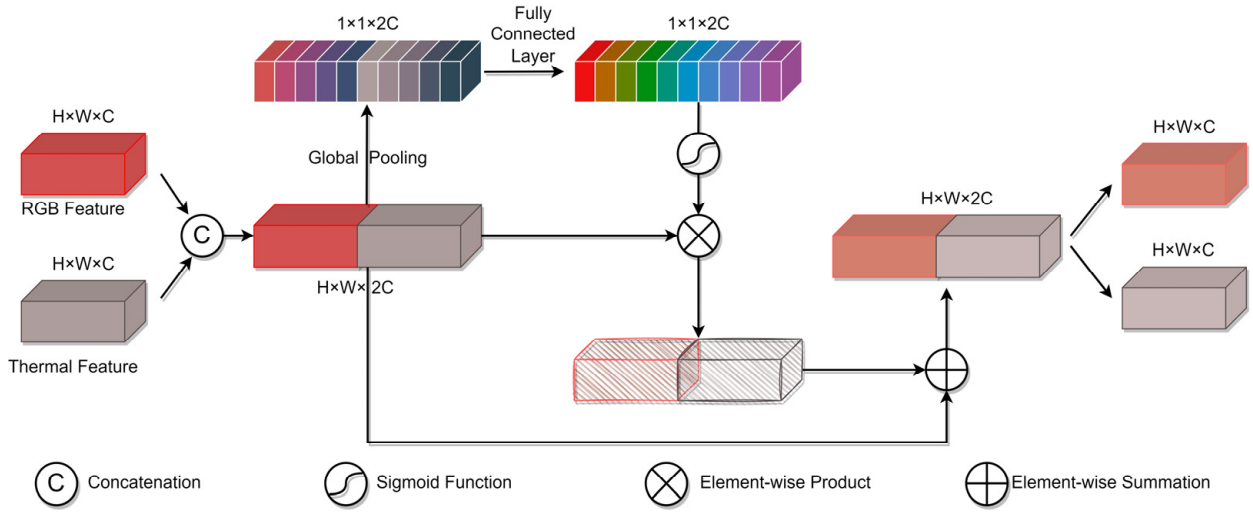


Figure 5. Illustration of RCAE. RCAE first concatenates the features extracted from visible light and thermal infrared images and obtains the weight vector through GAP. Then, the new feature is obtained using the residual connection. Finally, RCAE separates the output according to the channel.

The enhanced visible light features $\varphi_r^*(z_r)$ and thermal infrared features $\varphi_t^*(z_t)$ are obtained by separating the output z according to the channel. The whole process of RCAE can be understood to learn the weight coefficient of each channel through channel self-attention, which enhances the features and representational power of the network.

3.4. Maximum Fusion Prediction Module

The region proposal network, as utilized in SiamRPN [20], comprises two branches: a classification branch for foreground and background classification and a regression branch for proposal regression. In cases where there are k anchors, the network is required to output $2k$ channels for classification and $4k$ channels for regression. In SiamCAF, the outputs of CCF and RCAE are fed into the region proposal subnetwork. Taking the output of CCF as an instance study, the feature maps z_f in classification branch require the expansion of the number of channels to $2k$ through the convolutional layer, while the corresponding feature maps x_f in the detection branch, as an input, require additional size transformation through the convolutional layer, without expanding the number of channels. We denote the operation through the convolutional layer as $(\cdot)_{cls}$. Thus, z_f and x_f, after undergoing the convolutional layer, can be denoted as $\left(z_f\right)_{cls}$ and $\left(x_f\right)_{cls}$, and the correlation calculation of the two features can be obtained:

$$A_{clsF} = \left(x_f\right)_{cls} \bigstar \left(z_f\right)_{cls} \tag{10}$$

where ★ denotes the correlation calculation, and the feature maps $\left(z_f\right)_{cls}$ are used as kernels. Similarly, the classification branch of the outputs of the RCAE can be obtained as follows:

$$A_{clsR} = (\varphi_r^*(x_r))_{cls} \bigstar (\varphi_r^*(z_r))_{cls} \tag{11}$$

$$A_{clsT} = (\varphi_t^*(x_t))_{cls} \bigstar (\varphi_t^*(z_t))_{cls} \tag{12}$$

The SoftMax loss is utilized to provide supervision for the classification branch. In MFP, we need to fuse the three predicted classifications in order to predict each anchor at the corresponding location on the original map. We incorporate scale change penalties for positive predictions to mitigate abrupt changes in size and aspect ratio. We introduce a cosine box to reduce the impact of large displacements in the same way as SiamRPN and then choose the largest item via combination:

$$S_r = Cos(penalty \times A_{clsR}) \tag{13}$$

$$S_t = Cos(penalty \times A_{clsT}) \tag{14}$$

$$S_f = Cos(penalty \times A_{clsF}) \tag{15}$$

$$S = max\left(S_f + S_t, S_f + S_r\right) \tag{16}$$

where S_r denotes the map of the positive prediction of the enhanced visible light features, S_t represents the prediction map of the enhanced thermal infrared features, and S_f represents the prediction map of the fused features. To comprehensively consider the varying capabilities of representation in different features, we use S to fuse the corresponding elements for obtaining a more accurate and reliable predicted position of the object.

4. Experiment and Result Analysis

This section begins with a comparison between our proposed SiamCAF and other advanced RGBT tracking methods to showcase its superior performance. Subsequently, we conduct ablation experiments to validate the effectiveness of each module and the different modalities in our approach. Finally, we discuss the implementation details.

4.1. Evaluation Dataset and Evaluation Metrics

4.1.1. GTOT Dataset and Metrics

The GTOT dataset [2] contains 50 video pairs in different scenes and conditions, with each pair consisting of a visible light video and a thermal infrared video. It consists of frames with artificially marked ground truth, and the challenge attributes are categorized into seven groups based on the state of the target. The videos in the dataset exhibit high diversity, and to ensure consistency in the annotations, they were all completed by a single person. Two widely used evaluation metrics, the precision rate (PR) and success rate (SR) in one-pass evaluation (OPE), are used as evaluation indicators of the tracker. For GTOT, where the target object is typically small, we set the threshold to five pixels following the previous work.

4.1.2. RGBT234 Dataset and Metrics

The RGBT234 dataset [3] is a large-scale RGBT tracking dataset. It is an expanded version of the RGBT210 dataset [33] and has 234 sequences and 12 challenge attributes. The acquisition equipment for this dataset is a thermal infrared camera and a CCD camera, and the imaging parameters of the two cameras are consistent, which can ensure that the alignment between the visible light and thermal infrared sequence pairs is highly accurate, and no preprocessing or post-processing is required. Following the previous work, our

evaluation metrics are based on the maximum precision rate (MPR) and the maximum success rate (MSR), and the threshold is 20 pixels. To be specific, we computed the PR/SR for both the visible light and thermal infrared modalities and then selected the maximum value between the two as our MPR/MSR.

4.1.3. VTUAV Dataset and Metrics

The VTUAV dataset [5] was captured using a professional UAV. It comprises a total of 500 sequences, containing 1,664,549 RGB-T image pairs. The dataset is split into two different sets: a training set comprising 250 sequences and a separate test set consisting of the remaining 250 sequences. Additionally, to account for the presence or absence of targets, all sequences are further categorized into long-term and short-term sets, allowing for a thorough evaluation of the tracking performance in different scenarios. In this paper, we focus on the short-term set, and its challenges are summarized as 13 attributes. In this evaluation, we continue to utilize the maximum precision rate (MPR) and maximum success rate (MSR) as quantitative measures to assess the performance, in the same way as for the RGBT234 dataset.

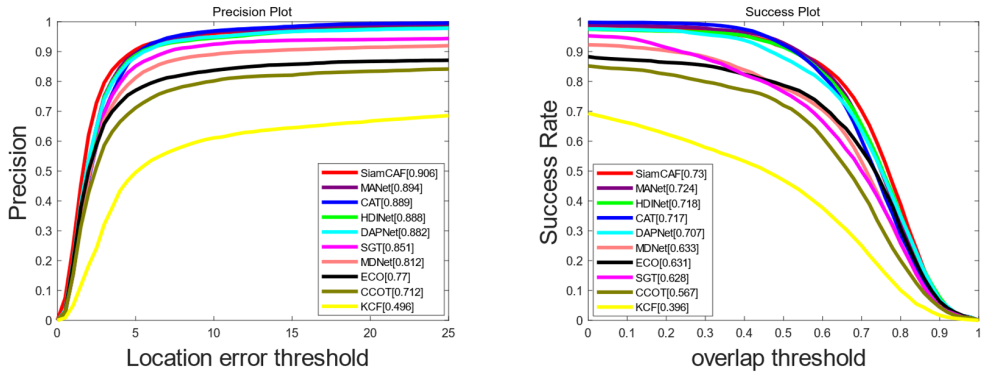
4.2. Implementation Details

The parameters of our backbone are initialized using the modified AlexNet, pretrained from ImageNet. In detail, the first three convolution layers are fixed, and we only fine-tune the last two convolution layers in SiamCAF. During the training process, we utilize the SGD optimizer with an initial learning rate of 10^{-2} and an end learning rate of 10^{-5}. Additionally, the momentum is 0.9 and weight decay is 5×10^{-4}. We set the batch size as 28 and train the model for 50 epochs in total. The template and the detection patches are extracted in the same way as SiamRPN. Our tracker is implemented in Python using PyTorch. All experiments are run with a NIVIDA GeForce RTX 3090 GPU and an Intel I9-10980XE CPU.

4.3. Result Comparisons on GTOT

4.3.1. Overall Performance

Based on the GTOT dataset, we compared SiamCAF with other advanced RGB trackers (KCF [17], ECO [34], C-COT [19], and MDNet [12]) and advanced fusion trackers (CAT [35], SGT [33], MANet [36], DAPNet [37], and HDINet [38]). According to Figure 6, our SiamCAF demonstrates a superior performance, achieving a success rate of 73% and a precision rate of 90.6%. Our method also achieves a clear improvement over the other RGB trackers, proving the importance of thermal information in object tracking. Notably, when compared with the most recent state-of-the-art tracker, HDINet, our algorithm exhibits improvements of 1.2% in the success rate and 1.8% in the precision rate. Furthermore, our SiamCAF demonstrates remarkable speed on the GTOT dataset, being 116 times faster than HDINet.



(**a**) (**b**)

Figure 6. Comparison of PR and SR on GTOT dataset: (**a**) precision plot of GTOT; (**b**) success plot of GTOT.

4.3.2. Attribute-Based Performance

We conducted a comprehensive comparison of SiamCAF with other advanced RGBT trackers on subsets that involved diverse challenge attributes, including KCF, SRDCF [39] +RGBT, RT-MDNet [40], DuSiamRT, SGT, MANet, DAPNet, and HDINet. The evaluation results are presented in Table 1, demonstrating that our SiamCAF method consistently outperforms the other RGBT trackers in the majority of the challenges, providing compelling evidence for the effectiveness of our approach.

Table 1. Attribute-based PR/SR scores (%) against other trackers on the GTOT dataset. The best and second-best results are presented in red and blue, respectively.

Attributes	OCC	LSV	FM	LI	TC	SO	DEF	ALL
KCF+RGBT	52.2/35.9	55.4/41.4	42.6/34.2	45.9/37.8	44.9/36.1	44.4/30.9	49.2/40.0	49.6/39.6
SRDCF+RGBT	72.7/58.0	80.4/68.1	68.3/61.1	71.7/59.4	70.5/58.0	80.5/57.5	66.6/53.7	71.9/59.1
RT-MDNet	73.3/57.6	79.1/63.7	78.1/64.1	77.2/63.8	73.7/59.0	85.6/63.4	73.1/61.0	74.5/61.3
DuSiamRT	72.8/57.7	80.9/64.5	72.1/58.0	76.2/62.3	78.1/61.4	84.4/64.2	76.4/62.9	76.6/62.8
SGT	81.0/56.7	82.6/55.7	82.0/55.7	84.3/59.0	84.4/59.6	85.7/60.0	86.7/62.1	85.1/62.8
DAPNet	87.3/67.4	84.7/66.1	82.3/65.3	90.0/67.7	89.3/68.0	93.7/68.2	91.9/69.6	88.2/70.7
HDINet	86.3/66.9	87.9/70.8	88.2/70.4	91.9/74.5	87.0/68.9	95.3/70.5	90.3/73.8	88.8/71.8
MANet	88.2/69.6	87.6/70.1	87.6/69.9	89.0/71.2	89.0/71.0	89.7/70.8	90.1/71.5	89.4/72.4
SiamCAF	89.8/69.9	88.0/69.5	88.0/68.6	91.8/74.4	90.1/71.6	91.7/70.5	92.2/75.4	90.6/73.0

4.3.3. Visual Comparison

We compared SiamCAF with five advanced trackers, namely, MANet + RGBT, KCF, DAPNet, SGT, and SiameseFC, on three sequences. As shown in Figure 7, SiamCAF could accurately track the target, while the most popular algorithms for comparison failed to track it. When the target was severely occluded or in thermal crossover, our tracker effectively handled this challenge with a high performance, because the CCF can fuse two modalities in a more superior manner. While the other trackers could lose track of the target when moving quickly and producing large-scale changes, our method maintained continuous tracking and copes with large-scale change throughout the video sequence because of the RCAE and RPN.

4.4. Result Comparisons on RGBT234

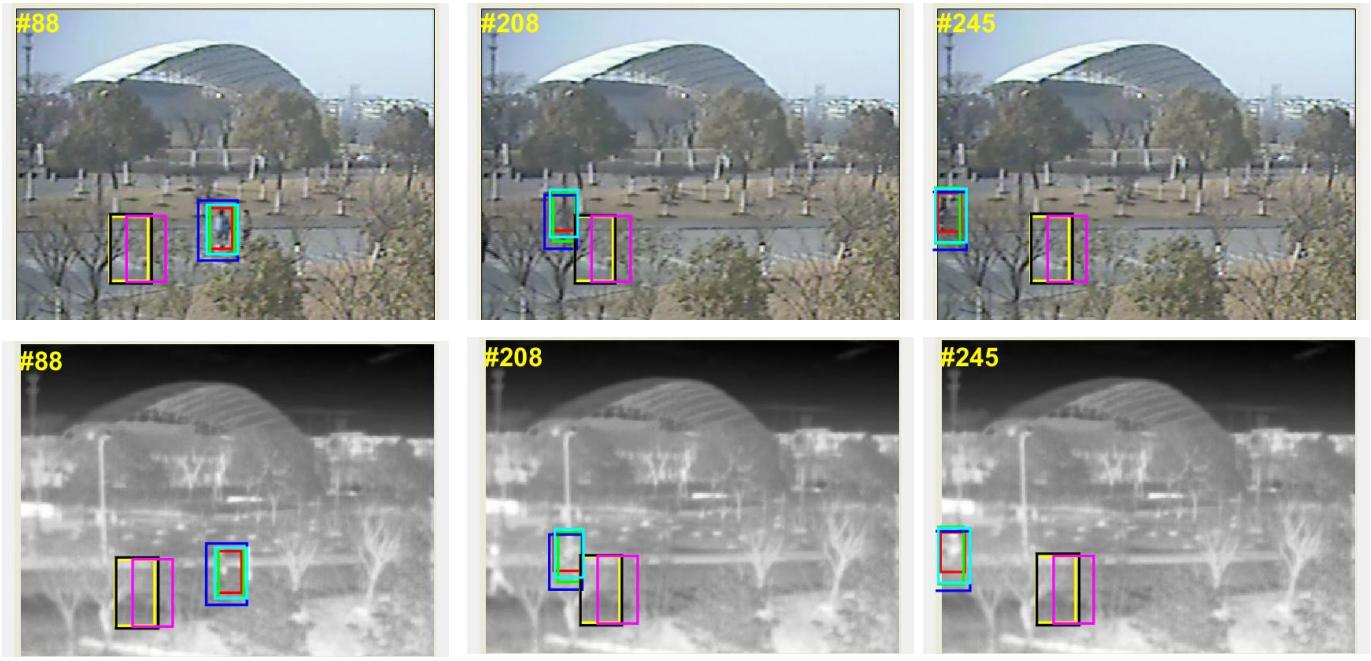
4.4.1. Overall Performance

On the RGBT234 dataset, we compared SiamCAF with other advanced RGB trackers (ECO [34] and C-COT [19]) and advanced fusion trackers (KCF [17] + RGBT, DAPNet [37], SGT [33], MDNet [12] + RGBT, SiamDW [41] + RGBT, CFnet [42] + RGBT, and SOWP [43] + RGBT). The results are shown in Figure 8. Our SiamCAF method realizes the best performance. On the RGBT234 dataset, the MPR/MSR score of SiamCAF reached 77.1%/53.7%. Specifically, SiamCAF scores 5.1% higher than SGT in the MPR, and 6.5% higher in the MSR, further proving the effectiveness of SiamCAF. Compared with DAPNet, our method has a greater advantage in the MPR, which may be due to the fact that we used the MFP in the classification branch, which makes the foreground–background classification more accurate.

4.4.2. Attribute-Based Performance

We conducted a comprehensive comparison of SiamCAF with other advanced RGBT trackers on subsets that involved 12 challenge attributes, including KCF + RGBT, DAPNet, SGT, MDNet + RGBT, SiamDW [41] + RGBT, CFnet [42] + RGBT, SOWP [43] + RGBT, L1-PF [44], and DSST [45]. The results of the evaluation are depicted in Figure 9. It is apparent that SiamCAF outperforms most of the trackers in all challenges. The attribute-based experiments clearly showcase the superior tracking capability of SiamCAF to effectively deal with a wide range of challenges.

(**a**)



(**b**)

Figure 7. *Cont.*

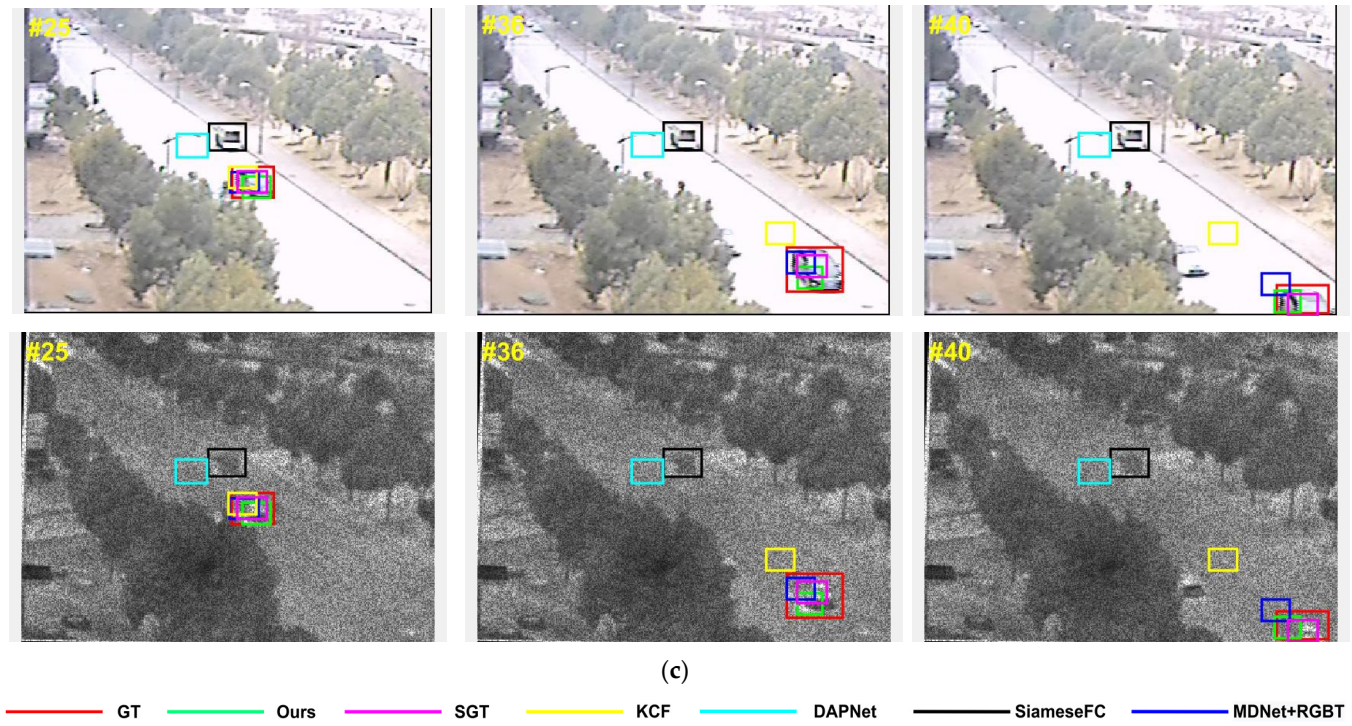


Figure 7. Qualitative comparisons with five popular trackers: (**a**) LightOcc; (**b**) occBik; (**c**) fastCar2.

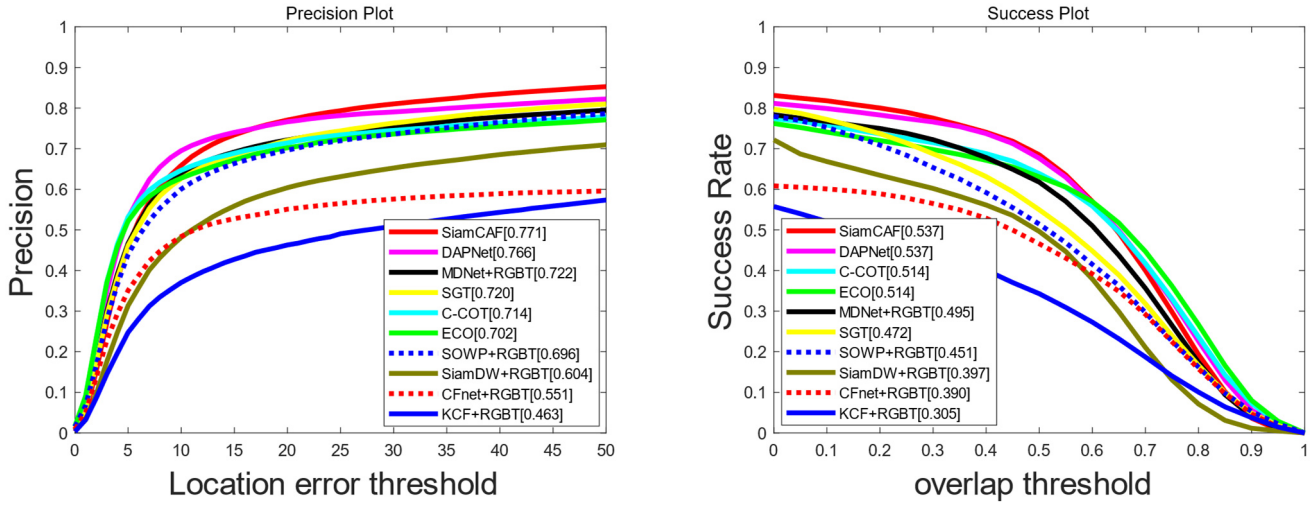
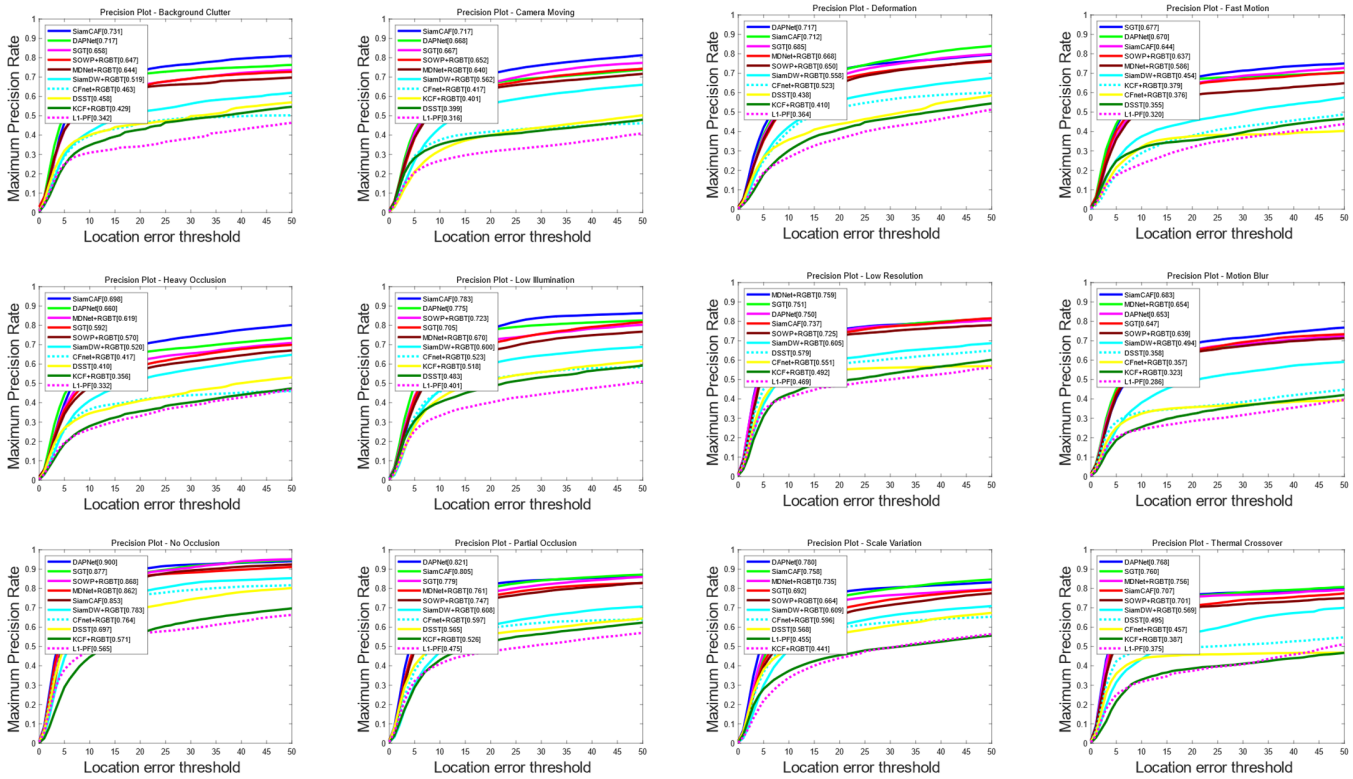


Figure 8. Comparison of MPR and MSR on the RGBT234 dataset: (**a**) precision plot of RGBT234; (**b**) success plot of RGBT234.

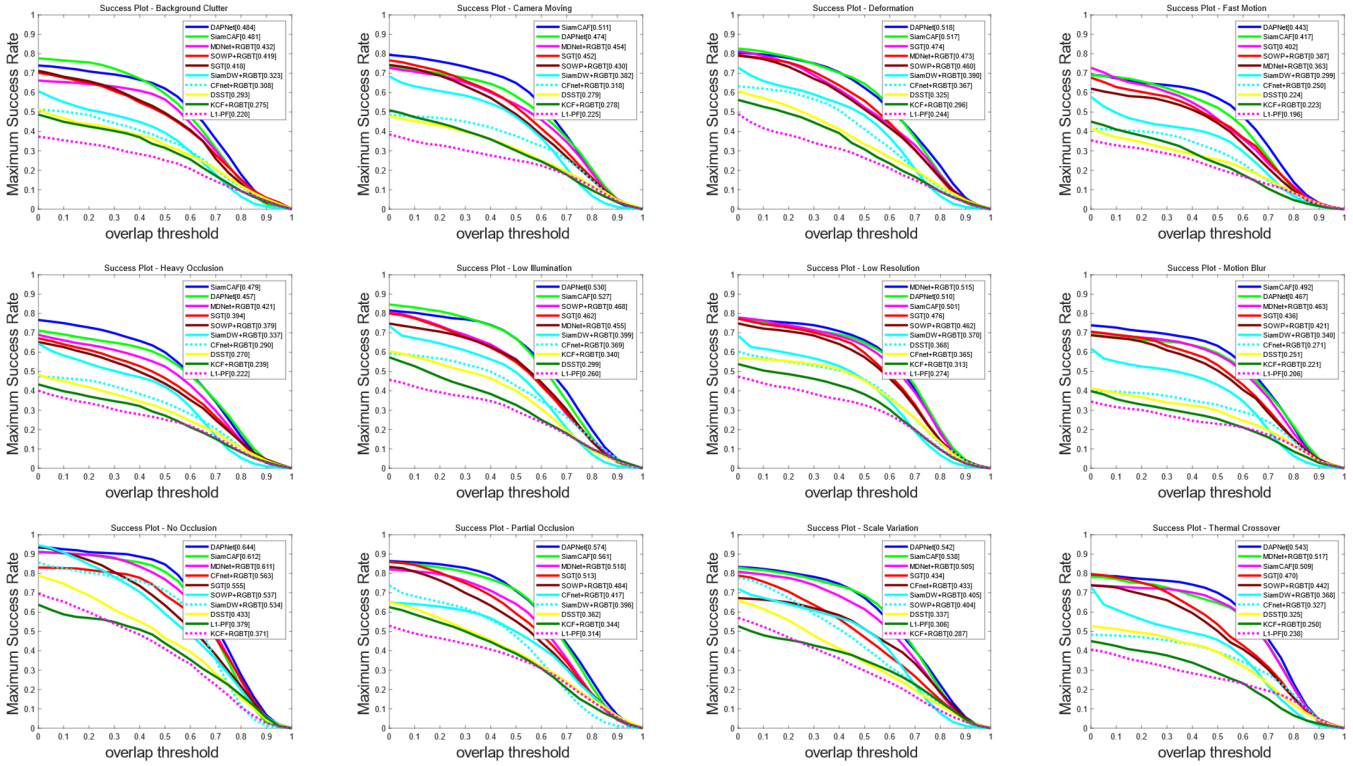
4.5. Result Comparisons on VTUAV

4.5.1. Overall Performance

Based on the VTUAV dataset, we compared SiamCAF with other advanced RGB trackers (GlobalTrack [46], SiamFC [13], SPLT [47], HiFT [48], D3S [49]) and state-of-the-art fusion trackers (DAFNet [50], ADRNet [51]). The evaluation results are presented in Figure 10. Our SiamCAF method realizes the optimal performance. On the VTUAV dataset, the MPR/MSR score of SiamCAF reached 67.0%/54.2%. Specifically, SiamCAF scores 8.9% higher than SiamFC in the MPR and 8.4% higher in the MSR, further proving the effectiveness of SiamCAF.

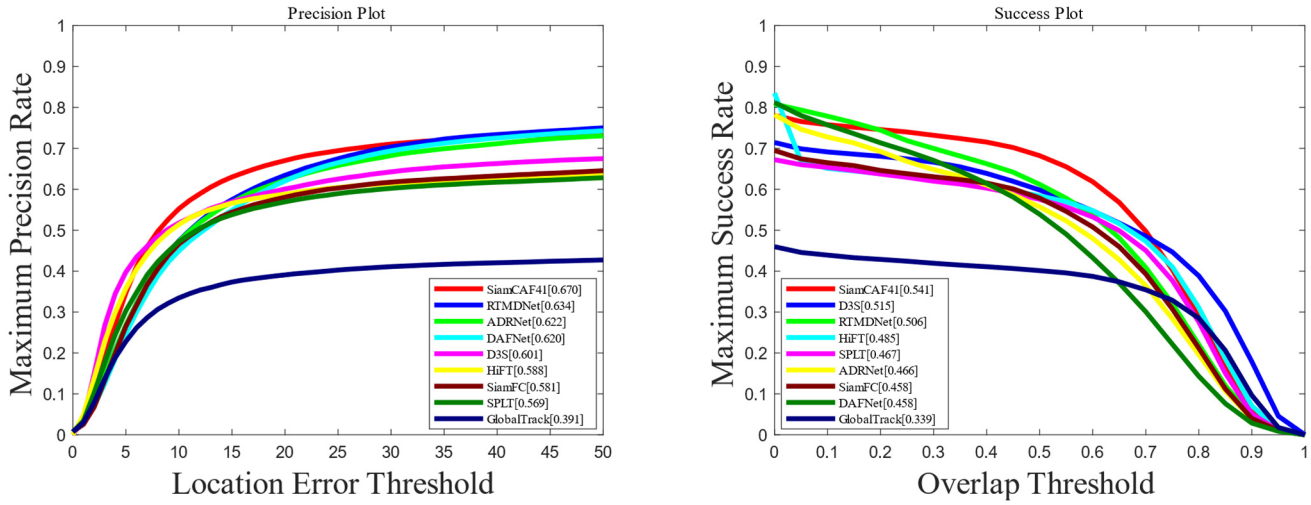


(**a**)



(**b**)

Figure 9. Comparison of MPR and MSR based on 12 attribute challenges: (**a**) evaluation of MPR curves; (**b**) evaluation of MSR.



(**a**) (**b**)

Figure 10. Comparison of MPR and MSR on the VTUAV dataset: (**a**) precision plot of VTUAV; (**b**) success plot of VTUAV.

4.5.2. Visual Comparison

As shown in Figure 11, we compared SiamCAF with five advanced trackers, namely, GlobalTrack, SiamFC, SPLT, HiFT, and DAFNet, on two sequences. In comparison to the other popular tracking algorithms, SiamCAF demonstrated superior performance in accurately tracking the target. Even in challenging scenarios such as fast movement or thermal crossover, our tracker was able to handle the challenge well due to the superior fusion capability of the CFF. Unlike other trackers that may lose track of the target when faced with large-scale changes, our method, which incorporates the use of the RCAE and RPN, maintains continuous tracking and effectively handles large-scale changes.

4.6. Ablation Study

We conducted the first ablation study on the GTOT dataset to validate the effectiveness of the key components of SiamCAF. Two degraded versions of SiamCAF were used, including SiamCAF-noRCAE, for which we removed the Residual Channel Attention Enhanced module, and SiamCAF-noCCF, in which the Complementary Coupling Feature fusion module was deleted. According to the experimental results in Figure 12, the following points could be obtained that SiamCAF scores 1.8%/1.1% higher than SiamCAF-noRCAE in the PR/SR, which shows that RCAE can amplify and enhance the characteristics of different modalities to achieve better tracking. The evaluation result of SiamCAF is 2.4%/3.0% greater than that of SiamCAF-noCCF, which verifies that CCF can extract the similar features and reduce the modality differences to better fuse the visible light and thermal infrared features. The experimental results confirm the feasibility of the main components of SiamCAF.

Furthermore, we conducted a second ablation experiment on both the GTOT and RGBT234 datasets to establish that the combination of visible light and thermal infrared modalities in SiamCAF yields a superior tracking performance compared to the use of a single modality. In the variant SiamCAF-RGB, only visible light sequences are input into the network, while in SiamCAF-T, only thermal infrared sequences are used. It can be seen from Figure 13 that SiamCAF scores 14%/9% higher than SiamCAF-RGB and 11.9%/7.8% higher than SiamCAF-T in terms of the objective evaluation indicators, the PR and SR, on GTOT. On RGBT234, SiamCAF scores 8.8%/6.3% higher than SiamCAF-RGB and 14.3%/14% higher than SiamCAF-T in the MPR/MSR. These results show that visible light and thermal infrared modalities can co-operate with each other to improve the accuracy of tracking, and SiamCAF accomplishes this task well.

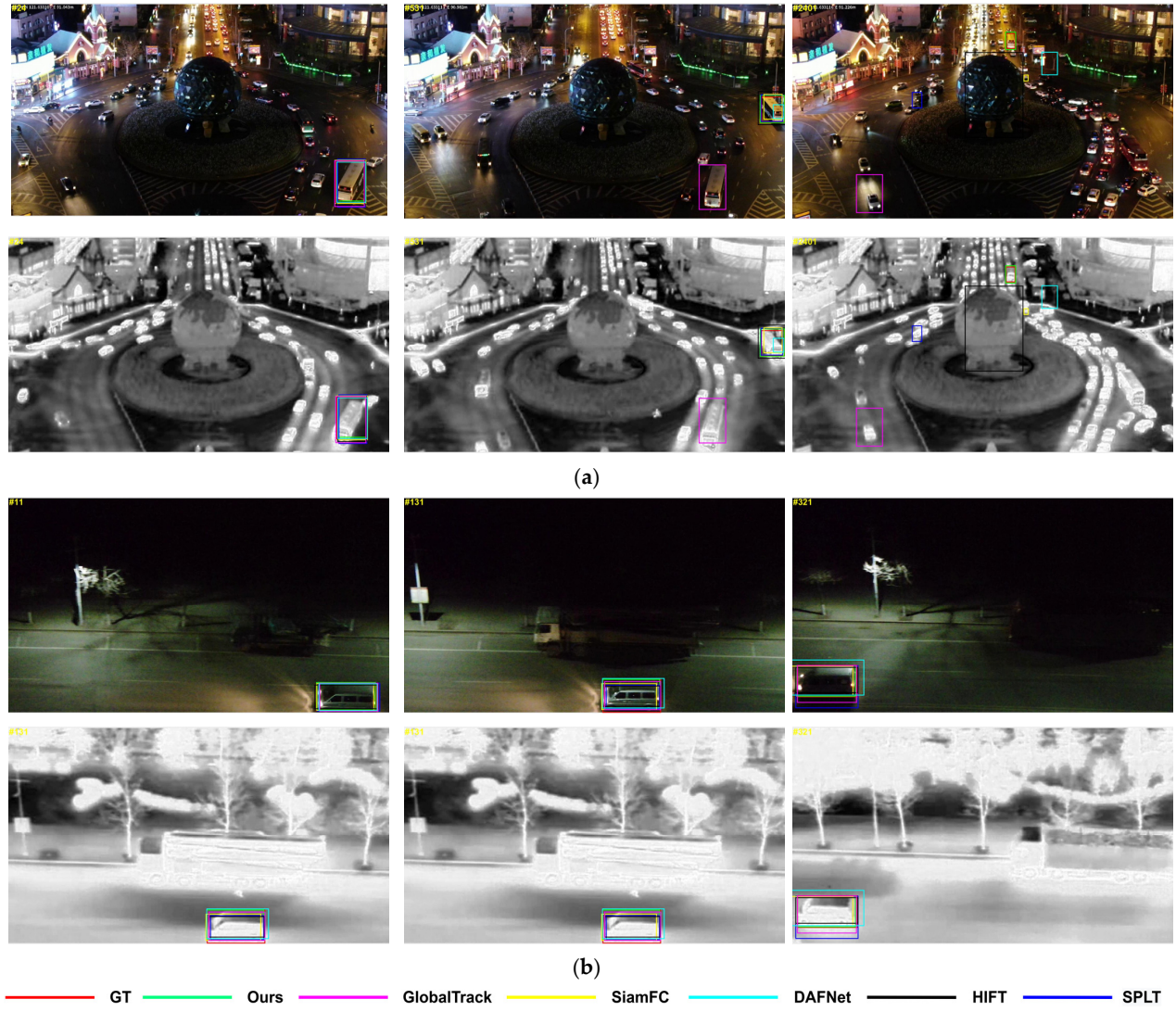
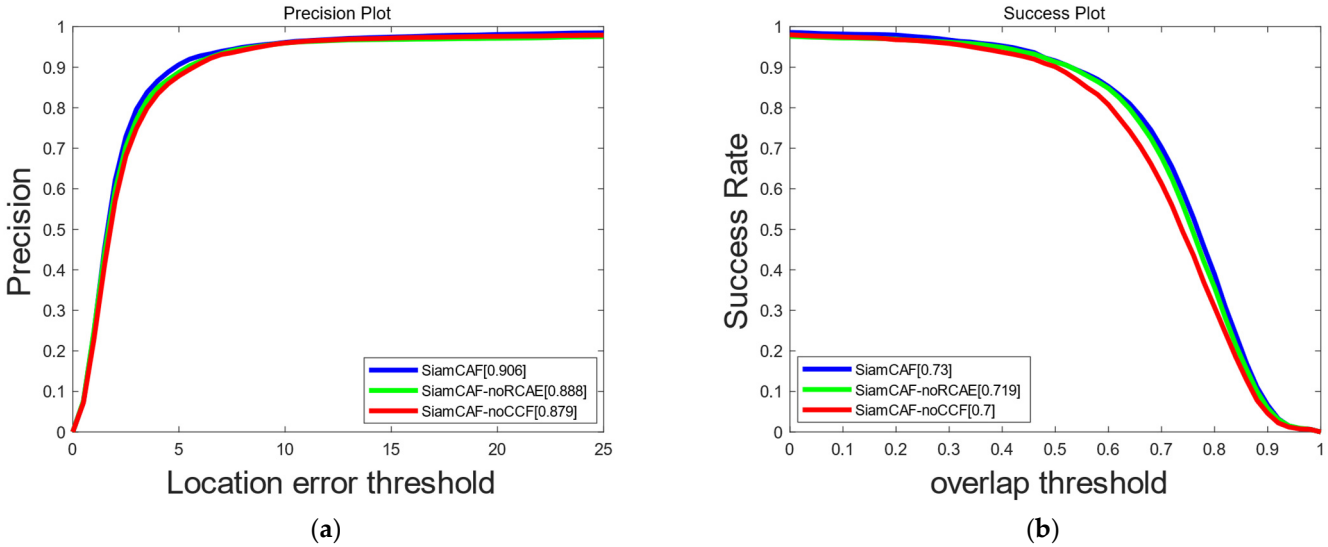


Figure 11. Qualitative comparisons with five popular trackers: (**a**) bus_014; (**b**) car_059.



(**a**) (**b**)

Figure 12. PR and SR of SiamCAF and the two variants. (**a**) Precision plot of GTOT. (**b**) Success plot of GTOT.

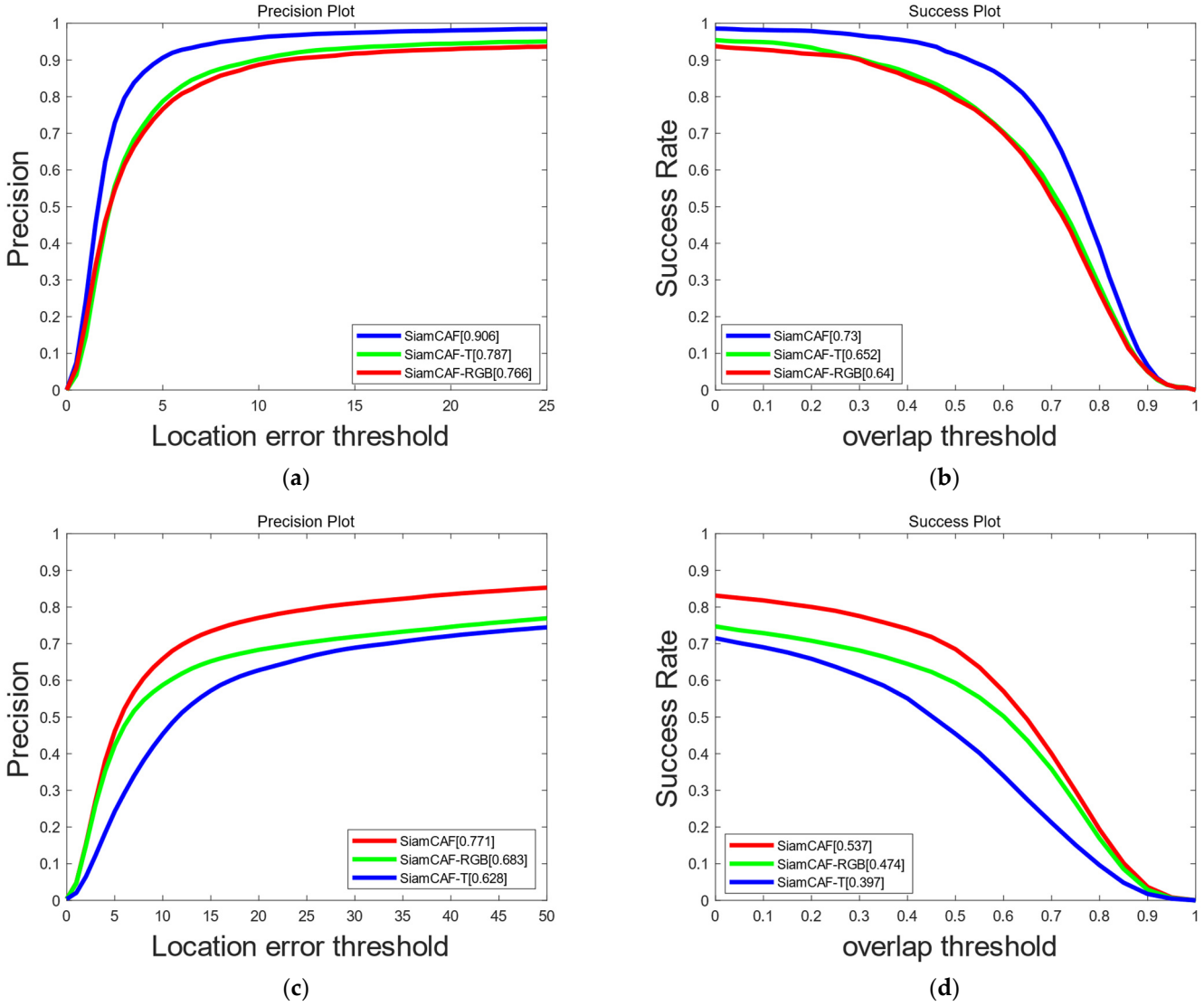


Figure 13. PR and SR of SiamCAF with different modalities. (**a**) Precision plot of GTOT. (**b**) Success plot of GTOT. (**c**) Precision plot of RGBT234. (**d**) Success plot of RGBT234.

4.7. Efficiency Analysis

We compared the efficiency of SiamCAF with that of other fusion tracking methods (SGT, mfDiMP [31], MANet [36], SiamDW [41] +RGBT, HDINet [38], DuSiamRT [14], and SiamFT [22]), as shown in Figure 14. It can be seen that the speed of the proposed SiamCAF greatly exceeds that of most of the fusion methods. SiamCAF reaches 105 FPS and has the best performance on the GTOT. SiamCAF balances robustness and speed at the same time. We used a dual-modal Siamese network to make the framework more concise. Simultaneously, RCAE and CCF are simpler and more convenient than the other fusion methods.

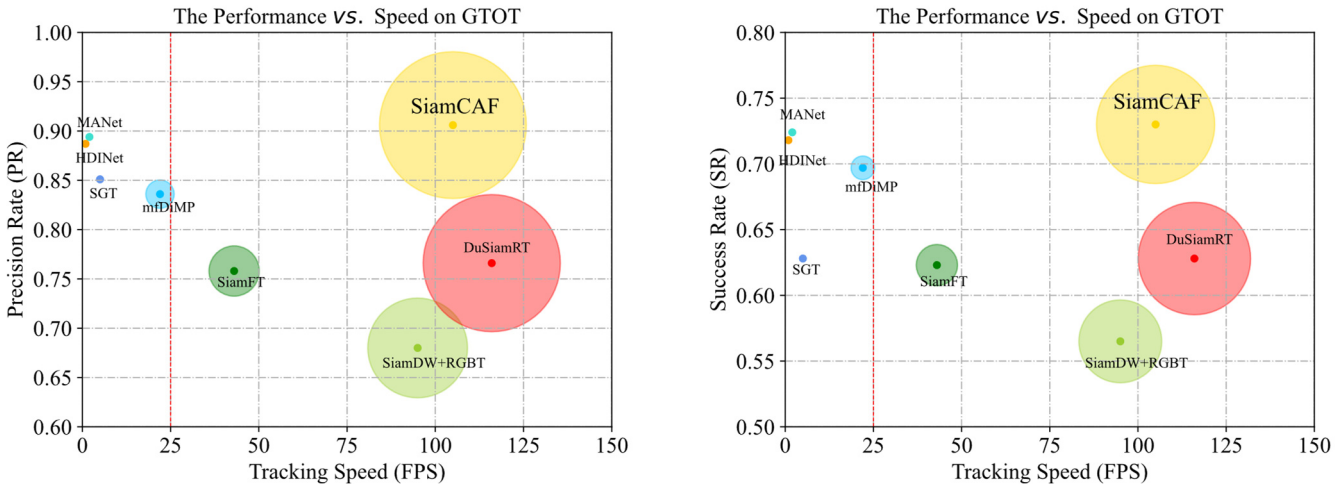


Figure 14. Speed comparison of various tracking methods: (**a**) PR and speed based on GTOT. (**b**) SR and speed based on GTOT.

5. Conclusions

A novel RGBT Siamese tracker called SiamCAF was proposed in this paper. By leveraging the collaborative power of newly designed modules, our method effectively exploits both visible light and thermal infrared features for RGBT tracking and achieves a state-of-the-art performance with a beyond-real-time running speed. In particular, due to the proposed CCF, our tracker can take full advantage of the complementary information of different modalities, and thus, satisfactory results were achieved in some challenging conditions, such as low illumination and heavy occlusion. Simultaneously, RCAE is designed to learn the weight coefficient of each channel through channel self-attention, which can enhance the features and representational power of the network. Finally, MFP completes the response-level fusion in the response map fusion stage. The extensive experimental results obtained on the GTOT, RGBT234, and VTUAV datasets demonstrate that our proposed SiamCAF tracker achieves a significantly improved performance compared to other state-of-the-art algorithms and can reach 105 FPS.

Author Contributions: Conceptualization, Y.X. and J.Z.; methodology, Y.X. and Z.L.; software, J.Z. and Y.Z.; validation, Y.X., J.Z. and C.L.; formal analysis, Y.Z., B.H. and C.L.; data curation, Y.X., Z.L. and B.H.; writing—original draft preparation, Y.X.; writing—review and editing, Y.X., J.Z. and Z.L.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (No. 62076137).

Data Availability Statement: All data included in this study are available upon request by contacting the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2013**, *25*, 245–262. [CrossRef]
2. Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; Lin, L. Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking. *IEEE Trans. Image Process.* **2016**, *25*, 5743–5756. [CrossRef] [PubMed]
3. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
4. Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; Sun, D. LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking. *IEEE Trans. Image Process.* **2022**, *31*, 392–404. [CrossRef] [PubMed]
5. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline. *arXiv* **2022**, arXiv:2204.04120.

6. Kulikov, G.Y.; Kulikova, M.V. The Accurate Continuous-Discrete Extended Kalman Filter for Radar Tracking. *IEEE Trans. Signal Process.* **2016**, *64*, 948–958. [CrossRef]
7. Changjiang, Y.; Duraiswami, R.; Davis, L. Fast multiple object tracking via a hierarchical particle filter. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 17–21 October 2005; pp. 212–219.
8. Li, Z.; Gao, J.; Tang, Q.; Sang, N. Improved mean shift algorithm for multiple occlusion target tracking. *Opt. Eng.* **2008**, *47*, 086402.
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
10. Cheung, W.; Hamarneh, G. n-SIFT: N-Dimensional Scale Invariant Feature Transform. *IEEE Trans. Image Process.* **2009**, *18*, 2012–2021. [CrossRef] [PubMed]
11. Jia, C.; Wang, Z.; Wu, X.; Cai, B.; Huang, Z.; Wang, G.; Zhang, T.; Tong, D. A Tracking-Learning-Detection (TLD) method with local binary pattern improved. In Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), Zhuhai, China, 6–9 December 2015; pp. 1625–1630.
12. Nam, H.; Han, B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4293–4302.
13. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision Workshop (ECCVW), Amsterdam, The Netherlands, 8–16 October 2016; pp. 850–865.
14. Guo, C.; Yang, D.; Li, C.; Song, P. Dual Siamese network for RGBT tracking via fusing predicted position maps. *Vis. Comput.* **2021**, *38*, 2555–2567. [CrossRef]
15. Bolme, D.S.; Beveridge, J.R.; Draper, B.A. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
16. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7575, pp. 702–715.
17. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 583–596. [CrossRef] [PubMed]
18. Danelljan, M.; Khan, F.S.; Felsberg, M.; Weijer, J.V.D. Adaptive Color Attributes for Real-Time Visual Tracking. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1090–1097.
19. Danelljan, M.; Robinson, A.; Shahbaz Khan, F.; Felsberg, M. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; Volume 9909, pp. 472–488.
20. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
21. Zhang, X.; Ye, P.; Qiao, D.; Zhao, J.; Peng, S.; Xiao, G. Object Fusion Tracking Based on Visible and Infrared Images Using Fully Convolutional Siamese Networks. In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
22. Zhang, X.; Ye, P.; Peng, S.; Liu, J.; Gong, K.; Xiao, G. SiamFT: An RGB-Infrared Fusion Tracking Method via Fully Convolutional Siamese Networks. *IEEE Access* **2019**, *7*, 122122–122133. [CrossRef]
23. Zhang, X.; Ye, P.; Peng, S.; Liu, J.; Xiao, G. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *Signal Process. Image Commun.* **2020**, *84*, 115756. [CrossRef]
24. Zhang, T.; Liu, X.; Zhang, Q.; Han, J. SiamCDA: Complementarity- and Distractor-Aware RGB-T Tracking Based on Siamese Network. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1403–1417. [CrossRef]
25. Feng, M.; Su, J. Learning reliable modal weight with transformer for robust RGBT tracking. *Knowl.-Based Syst.* **2022**, *249*, 108945. [CrossRef]
26. Peng, J.; Zhao, H.; Hu, Z.; Zhuang, Y.; Wang, B. Siamese infrared and visible light fusion network for RGB-T tracking. *Int. J. Mach. Learn. Cybern.* **2023**. [CrossRef]
27. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.

31. Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; Weijer, J.v.d.; Khan, F.S. Multi-Modal Fusion for End-to-End RGB-T Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 2252–2261.
32. Li, Y.; Zhao, H.; Hu, Z.; Wang, Q.; Chen, Y. IVFuseNet: Fusion of infrared and visible light images for depth prediction. *Inf. Fusion* **2020**, *58*, 1–12. [CrossRef]
33. Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; Tang, J. Weighted Sparse Representation Regularized Graph Learning for RGB-T Object Tracking. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1856–1864.
34. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939.
35. Li, C.; Liu, L.; Lu, A.; Ji, Q.; Tang, J. Challenge-Aware RGBT Tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; Volume 12367, pp. 222–237.
36. Li, C.L.; Lu, A.; Zheng, A.H.; Tu, Z.; Tang, J. Multi-Adapter RGBT Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 2262–2270.
37. Zhu, Y.; Li, C.; Luo, B.; Tang, J.; Wang, X. Dense Feature Aggregation and Pruning for RGBT Tracking. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 465–472.
38. Mei, J.; Zhou, D.; Cao, J.; Nie, R.; Guo, Y. HDINet: Hierarchical Dual-Sensor Interaction Network for RGBT Tracking. *IEEE Sens. J.* **2021**, *21*, 16915–16926. [CrossRef]
39. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
40. Jung, I.; Son, J.; Baek, M.; Han, B. Real-Time MDNet. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 83–98.
41. Zhang, Z.; Peng, H. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4586–4595.
42. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H.S. End-to-End Representation Learning for Correlation Filter Based Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5000–5008.
43. Kim, H.U.; Lee, D.Y.; Sim, J.Y.; Kim, C.S. SOWP: Spatially Ordered and Weighted Patch Descriptor for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3011–3019.
44. Wu, Y.; Blasch, E.; Chen, G.; Bai, L.; Ling, H. Multiple source data fusion via sparse representation for robust visual tracking. In Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 5–8 July 2011; pp. 1–8.
45. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; BMVA Press: London, UK, 2014.
46. Huang, L.; Zhao, X.; Huang, K. GlobalTrack: A Simple and Strong Baseline for Long-Term Tracking. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11037–11044. [CrossRef]
47. Yan, B.; Zhao, H.; Wang, D.; Lu, H.; Yang, X. 'Skimming-Perusal' Tracking: A Framework for Real-Time and Robust Long-Term Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2385–2393.
48. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. HiFT: Hierarchical Feature Transformer for Aerial Tracking. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15437–15446.
49. Lukežič, A.; Matas, J.; Kristan, M. D3S—A Discriminative Single Shot Segmentation Tracker. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7131–7140.
50. Gao, Y.; Li, C.; Zhu, Y.; Tang, J.; He, T.; Wang, F. Deep Adaptive Fusion Network for High Performance RGBT Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October; pp. 91–99.
51. Zhang, P.; Wang, D.; Lu, H.; Yang, X. Learning Adaptive Attribute-Driven Representation for Real-Time RGB-T Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 2714–2729. [CrossRef]