sensors

MDPI

Article

Multi-Scale Feature Interactive Fusion Network for RGBT Tracking

Xianbing Xiao [1], Xingzhong Xiong [2,*], Fanqin Meng [2] and Zhen Chen [1]

[1] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China
[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China
* Correspondence: xzxiong@suse.edu.cn

Abstract: The fusion tracking of RGB and thermal infrared image (RGBT) is paid wide attention to due to their complementary advantages. Currently, most algorithms obtain modality weights through attention mechanisms to integrate multi-modalities information. They do not fully exploit the multi-scale information and ignore the rich contextual information among features, which limits the tracking performance to some extent. To solve this problem, this work proposes a new multi-scale feature interactive fusion network (MSIFNet) for RGBT tracking. Specifically, we use different convolution branches for multi-scale feature extraction and aggregate them through the feature selection module adaptively. At the same time, a Transformer interactive fusion module is proposed to build long-distance dependencies and enhance semantic representation further. Finally, a global feature fusion module is designed to adjust the global information adaptively. Numerous experiments on publicly available GTOT, RGBT234, and LasHeR datasets show that our algorithm outperforms the current mainstream tracking algorithms.

Keywords: RGBT tracking; multi-scale feature; information interaction; transformer; attention mechanism

Citation: Xiao, X.; Xiong, X.; Meng, F.; Chen, Z. Multi-Scale Feature Interactive Fusion Network for RGBT Tracking. *Sensors* **2023**, *23*, 3410. https://doi.org/10.3390/s23073410

Academic Editors: Bingbing Gao and Gaoge Hu

Received: 23 February 2023
Revised: 16 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

1. Introduction

Visual object tracking [1,2] is widely used in autonomous driving, medical diagnostics, traffic monitoring, and other fields. Visual tracking aims at estimating the position and scale of the object in subsequent frames, given its state in the first frame. Although many RGB-based tracking algorithms have achieved excellent results, the performance of these trackers still needs to be further improved in some challenging scenarios. For example, the poor imaging quality of RGB in severe weather conditions usually leads to tracking failures. Therefore, RGBT tracking, which can make use of the advantages of RGB and infrared, has received extensive attention. Specifically, the infrared image is formed by thermal radiation with the benefits of insensitivity to light and intense penetration [3]. RGB images have rich texture and color information. Therefore, RGBT tracking can better handle complex scenarios due to its complementary advantages. Complementary advantages are shown in Figure 1.

To obtain more accurate and robust target attribute information and compensate for the information uncertainty of a single modality in target tracking, it is necessary to fuse the data collected by multi-sensors. The existing fusion tracking methods can be divided into pixel-level, feature-level, and decision-level. Pixel-level fusion tracking means that heterogeneous images are fused first, then target tracking is carried out based on the rich images [4,5]. Although rich information is preserved, it will bring greater computational costs during tracking. In feature-level fusion tracking, the features of RGB and infrared images are first extracted and then fused according to the designed fusion rules to obtain the fused feature and finally use the fused feature to perform tracking [6–8]. Decision-level fusion tracking is first performed in individual modalities to obtain tracking results or response maps. The results or response maps are then fused to get the final tracking

result [9,10]. The computational cost is relatively low compared to pixel-level and feature-level fusion tracking methods.

Figure 1. Illustration of the advantages of RGB and infrared modalities. (**a**) The advantage of thermal infrared modality over RGB modality. (**b**) The advantage of RGB modality over thermal infrared modality.

Although RGBT tracking has made significant progress, the mainstream methods can be roughly divided into two categories, namely the combined fusion tracking method and the discriminant fusion tracking method. The combined fusion tracking method aims to mine all valuable information of different modalities and combine them to achieve a richer representation. Li et al. [11] proposed a multi-adapter network to exploit modality-shared and modality-specific features, composed of a generic adapter, a modality adapter, and an instance adapter. Li et al. [12] proposed a challenge-aware network, summarizing existing datasets into several challenge attributes and training challenge branches one by one, which effectively enhanced the feature discrimination ability of poor modality. Although combined fusion is an effective fusion method, it tends to introduce noise and redundant information. However, the discriminant fusion tracking method aims to obtain discriminant information and achieve an effective fusion of information. To eliminate redundant information, Zhu et al. [13] proposed a fusion method based on the pruning strategy, using global average pooling and weighted random selection operations to score each channel and select high-quality channels for tracking. Zhang et al. [6] used the attention mechanism to calculate the reliability weights of multi-layer features to fuse multi-level deep features. Xu et al. [14] proposed a multi-modalities cross-layer bilinear pooling network, which uses channel and spatial attention mechanisms to predict the reliable weight of each position. However, these algorithms do not fully mine multi-scale information, limiting tracking performance when the target scale varies greatly. Moreover, they ignore the rich contextual information between features, and the interactive fusion of features can achieve more robust representations.

In response to the above problems, inspired by inception modules [15–17], we proposed a multi-scale feature interactive fusion network for RGBT tracking. Specifically, we add multiple branches to each layer for multi-scale feature extraction. In order to achieve effective fusion, we design a feature selection module that uses the channel-aware mechanism to calculate reliability weights for each branch and adaptively aggregate features from multiple branches. At the same time, the encoder and decoder of the Transformer are used to achieve self-enhancement and interaction enhancement. In addition, we designed a global feature fusion module to balance the contribution of different modalities in different scenarios, which adaptively adjusts global features in spatial and channel dimensions.

Finally, three fully connected layers are utilized for instance learning and target state estimation.

To sum up, the main contributions of this work are as follows:

i. We propose a new multi-scale feature interactive fusion network (MSIFNet) to implement robust RGBT tracking. The network can improve the recognition ability of targets of different sizes by fully exploiting multi-scale information, thus improving tracking accuracy and robustness;
ii. We design a feature selection module that adaptively selects multi-scale features for fusion by the channel-aware mechanism while effectively suppressing noise and redundant information brought by multiple branches;
iii. We propose a Transformer interactive fusion module to further enhance the aggregated feature and modality-specific features. It improves long-distance feature association and enhances semantic representation by exploring rich contextual information between features;
iv. We design a global feature fusion module, which adaptively adjusts the global information in spatial and channel dimensions, respectively, to integrate the global information more effectively.

2. Related Work

Visual tracking is a basic computer vision task, and many excellent algorithms have been proposed. This chapter will introduce relevant work from the following two aspects: (1) RGB tracker and (2) RGBT tracker.

(1) RGB trackers

Based on the correlation filter tracking algorithm [18,19], it has attracted ascendant attention because of its real-time speed, but handcrafted features limit the recognition ability. In response to this problem, siamFC [20] adopts the Siamese structure to introduce deep learning into tracking. With the application of Siamese networks in the tracking field, RGB tracking algorithms have developed rapidly. Additionally, a large number of algorithms based on Siamese networks have emerged, for example, siamRPN [21], SiamRPN++ [22] based on anchor, and siamFC++ [23], siamBAN [24], SiamCAR [25] based on the anchor-free mechanism. To optimize the tracking algorithm, Danelljan et al. [26] used off-line training IoU predictor and online training classifier for target state estimation, and they performed classification and regression tasks simultaneously in the tracking process to achieve robust tracking performance. To take advantage of the powerful feature expression ability of the Transformer, Wang et al. [27] used the Transformer to construct temporal information at different moments to model global dependencies better. Mayer et al. [28] aimed at the optimization-based tracking method that limits the expressive ability of the tracking network, used the Transformer to capture global relationships, and learned a more robust target prediction model.

(2) RGBT trackers

Traditional-based methods: Traditional RGBT tracking methods are mainly divided based on sparse representation and graph models. Since sparse representation can suppress noise and errors, early methods mainly focus on sparse representation. For example, Wu et al. [29] applied sparse representation to RGBT tracking for the first time, which integrated image patches of different modalities into a one-dimensional vector and carried out sparse representation in the target template. Li et al. [30] designed a fusion method based on Bayesian filtering, which considers intra-modality and inter-modality constraints for cross-modality sparse representation. Lan et al. [31] designed a modality-correlation-aware sparse representation model, adaptively selecting representative particles via low-rank and sparse regularization to handle appearance variations. The graph-based method can suppress the influence of noise background and has also received certain attention. To consider the synergism and heterogeneity between modalities, Li et al. [32] designed the cross-modality sorting graph model, introducing the cross-modality soft consistency

to integrate multi-modalities information effectively. To eliminate background clutter, Shen et al. [33] proposed a cooperative low-rank graph model, which decomposes the input features into low-rank feature parts and sparse noise parts and uses the cooperative graph learning algorithm to renew dynamically. However, these works utilize handcrafted features for tracking, which limits the performance to handle various challenges.

Deep learning-based methods: Deep learning is known for its robust feature expressiveness, which can model the appearance of objects better than hand-crafted methods. Xu et al. [34] first applied deep learning to the field of RGBT tracking. Subsequently, deep RGBT trackers have dominated. DAPNet [13] performs fusion at different feature levels using the same aggregation network. DAFNet [35] also adopts a similar strategy to DAPNet for feature fusion. CMPP [36] utilizes the attention mechanism to model correlations between heterogeneous data. Considering the importance of time information in the video sequence, the time context information is correlated to achieve more effective information inheritance. CAT [12], ADRNet figure [37], and APFNet [38] all model robust appearance representations by training multiple challenge-aware branches independently. However, the implementation details are different, CAT [12] adaptively aggregates multiple challenge branches and then adds them to the backbone feature learning process. ADRNet designs an attribute-driven residual branch that models different challenge attributes and aggregates them through residual connections to obtain a powerful target representation. In APFNet, different challenge branches are aggregated by SKNet [39] adaptively and use Transformers to enhance features.

The above trackers perform well, but these networks do not explore multi-scale information, limiting tracking performance when the target scale varies greatly. To solve the above problems, we design a new multi-scale feature interactive fusion network to handle the RGBT tracking task.

3. Methods

This section will introduce the proposed multi-scale feature interactive fusion network. First, we outline the entire architecture, and then we present the structure of the individual module.

3.1. Network Architecture

As shown in Figure 2, our network uses a symmetrical parallel structure to mine the potential information of two modalities. Specifically, we used the VGG-M [40] as the backbone feature extraction network. To trade off speed and accuracy, we used only the first 3 layers, with convolution kernel sizes of 7×7, 5×5, and 3×3. Following the first and second layer convolution are the Relu function, local response normalization, and max-pooling. After the third layer convolution, only the Relu function is used.

Different convolution kernels have various receptive fields, and different receptive fields can better adapt to targets of different sizes. Based on this knowledge, we add 5×5 and 3×3 convolution branches in the first layer to extract multi-scale features and add 3×3 and 1×1 convolution branches in the second layer. Considering the third layer of the backbone network is a small convolution kernel, only a 1×1 convolution branch is added to the third layer. For feature maps of different sizes, we use max-pooling to fix feature maps to the exact resolution for better fusion. To better integrate multi-scale features, a feature selection module was designed to adaptively activate features from different branches for aggregation while eliminating noise and redundancy. Then, the aggregated feature and modality-specific features are fed into the Transformer interactive fusion module for interaction enhancement. After fusion, the enhanced feature is added to backbone features for subsequent feature extraction. Furthermore, we designed a gated network to block noise propagation. Finally, the global feature fusion module is used to further balance the contributions of different modalities in different scenarios, and three fully connected layers and a softmax layer are used to predict the position of the target.

In MANet [11], modality-shared and modality-specific features are helpful for modeling target information. To take full advantage of shared and specific features, our backbone network does not share parameters to extract modality-specific features, while other pairs of convolutions (e.g., 2 convolution branches of 5×5 in the first layer) share parameters to extract shared information.

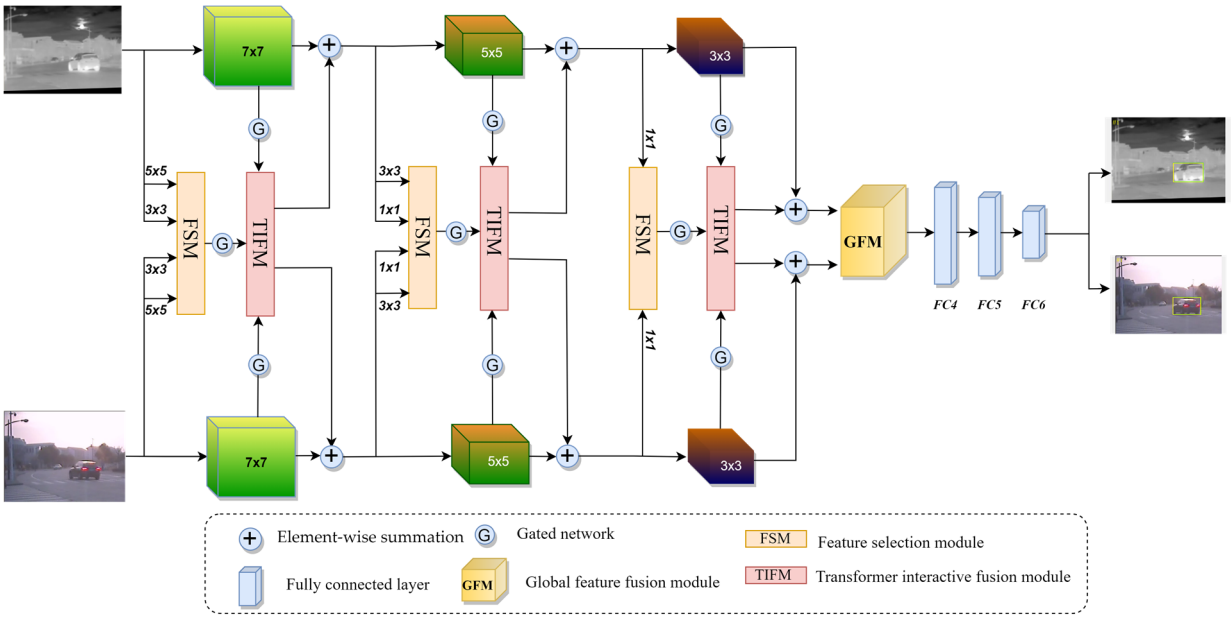

Figure 2. The overview of the proposed MSIFNet framework.

3.2. Feature Selection Module

There is redundant information in the feature maps of multiple branches. The reliability of features should be considered before fusion. Inspired by SKNet [39], we designed the feature selection module (FSM) to get more feature clues by deploying a small number of parameters. The FSM can adaptively activate valuable features for fusion while effectively suppressing noise and redundant information. Specifically, the FSM first performs simple element-wise summation, then uses the global average pooling operation (GAP) to obtain the channel vector. Then, feed the channel vector into 2 convolutions of 1×1 and a softmax function to obtain the channel weights. Finally, the original features are weighted with channel weights. The FSM adaptively activates features from different branches, combining them to focus the network on beneficial information. The specific situation of feature selection is shown in Figure 3. The whole process is summarized as follows:

$$X^a = \begin{cases} X^i_{5\times5} \oplus X^v_{5\times5} \oplus X^i_{3\times3} \oplus X^v_{3\times3}, & l = 1 \\ X^i_{1\times1} \oplus X^v_{1\times1} \oplus X^i_{3\times3} \oplus X^v_{3\times3}, & l = 2 \\ X^i_{1\times1} \oplus X^v_{1\times1}, & l = 3 \end{cases} \tag{1}$$

$$w_i = \phi\langle\delta\langle f(GAP(X^a))\rangle\rangle \tag{2}$$

$$X^{sel} = \begin{cases} w^i_{5\times5} \odot X^i_{5\times5} + w^v_{5\times5} \odot X^v_{5\times5} + w^i_{3\times3} \odot X^i_{3\times3} + w^v_{3\times3} \odot X^v_{3\times3}, & l = 1 \\ w^i_{3\times3} \odot X^i_{3\times3} + w^v_{3\times3} \odot X^v_{3\times3} + w^i_{1\times1} \odot X^i_{1\times1} + w^v_{1\times1} \odot X^v_{1\times1}, & l = 2 \\ w^i_{1\times1} \odot X^i_{1\times1} + w^v_{1\times1} \odot X^v_{1\times1}, & l = 3 \end{cases} \tag{3}$$

where l represents the l-th layer. f denotes 2 convolution operations of 1×1 inlaid with the ReLU function, expressed as $f = f_2(\xi(f_1))$. f_1 and f_2 represent 1×1 convolution operations. ξ and δ denote the ReLU function and softmax function, respectively. ϕ represents the reshape operation. w_i represents the generated weight. $\odot$ and $\oplus$ denote the element-wise product and element-wise summation.

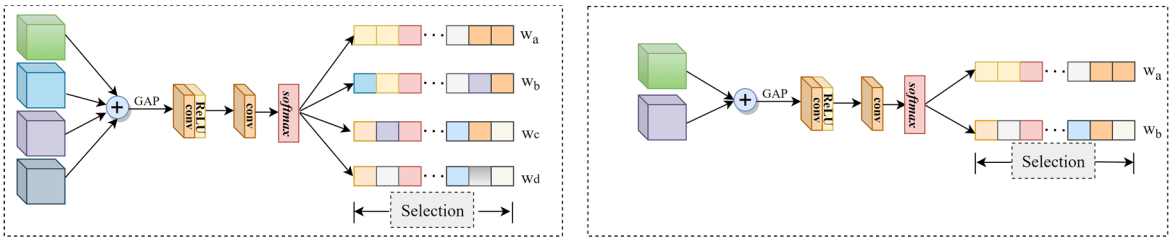

Figure 3. Specific structure of the FSM. The structure of the FSM of the first and second layers is shown in (**a**), and the third layer is shown in (**b**).

3.3. Transformer Interactive Fusion Module

The superior performance of the Transformer in computer vision proves the importance of modeling global dependencies. Inspired by [27,41], we present a Transformer interaction fusion module (TIFM) to explore the rich contextual information between features. The input of TIFM includes the modality-specific features and the aggregated feature (the output of the FSM). It should be noted that the aggregated feature and modality-specific features contain similar information (such as background information), and reducing redundant information can improve fusion efficiency. To solve this problem, a 1×1 convolution layer and sigmoid activation function are used to form the gated network to adjust the input adaptively.

As shown in Figure 4, to reduce the complexity of the model, we remove the position encoding and feed-forward network of the original Transformer. Each input feature is linearly transformed to produce a query matrix Q, a key matrix K, and a value matrix V. Through Q, K generates attention weights to modulate V. Q is then added to the output of Attention as a residual. Attention part can be expressed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{4}$$

where C is the dimension of key matrix to normalize attention. As we all know, the product of Q and K transpose can be regarded as the inner product of the corresponding vector so that the dependencies between any two elements of the global context can be modeled. Therefore, the encoders highlight the critical information by modeling its element dependencies. The decoders learn long-range dependencies among features to further enhance semantic information.

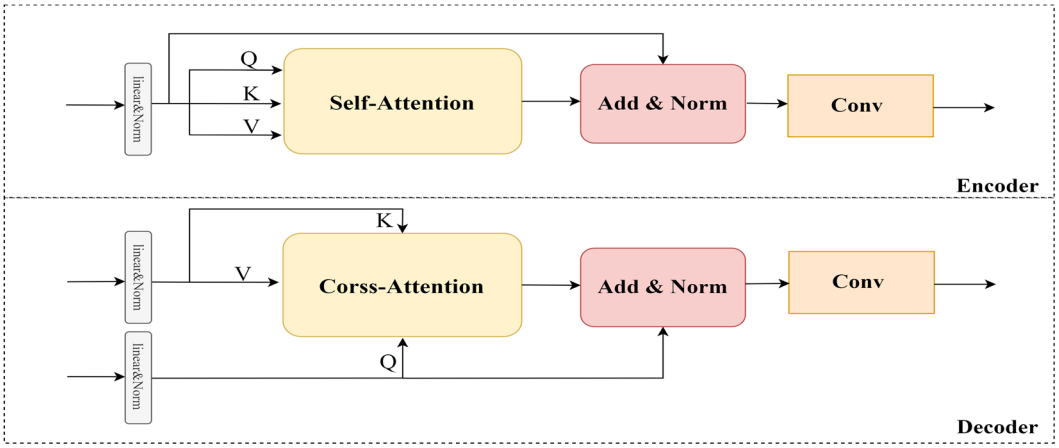

Figure 4. Specific structure of the encoder and decoder.

As shown in Figure 5, TIFM includes two encoders and two decoders. The encoders perform self-enhancement, and the decoders realize the interactive enhancement of the aggregated features (X^{sel}) and modality-specific features (X^{vis} and X^{inf}). Then, the element-wise summation operation combines the multiple enhancement features into a unified feature, and an encoder is connected to further enhance the semantic information. Specifically, we use the aggregated feature as the query for all features, resulting in a self-enhancement module (*SE*) and two feature interactive enhancement modules (*IE*). The whole process can be expressed as follows:

$$X^{\Theta} = SE\left(SE\left(X^{sel}, X^{sel}\right) + IE\left(X^{sel}, X^{vis}\right) + IE\left(X^{sel}, X^{inf}\right)\right) \tag{5}$$

where X^{Θ} represents the interactive enhanced features. Where $X^{sel}, X^{vis}, X^{inf} \in \mathbb{R}^{C\times H\times W}$ are the three inputs of TIFM. C, H, and W represent the number of channels, height, and width of the feature matrix.

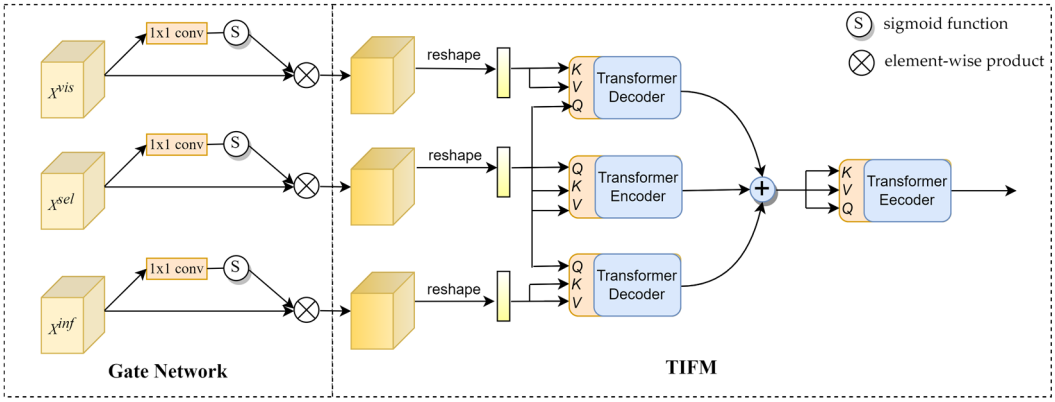

Figure 5. Specific structure of the Gate Network and TIFM.

3.4. Global Feature Fusion Module

In RGBT tracking, many algorithms aggregate the output features in the way of concatenating, which ignores the different contributions of different modalities in various scenarios. To weigh the contribution of the two modalities, we designed the global feature fusion module (GFM). The GFM aggregates multi-modalities features at the channel and spatial levels. As shown in Figure 6, it includes two branches, namely the channel integration branch and the spatial integration branch.

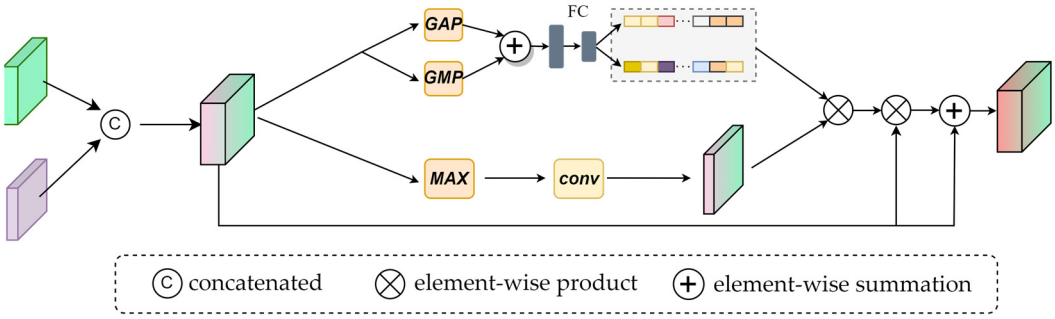

Figure 6. Specific structure of the GFM.

A channel integration branch similar to FSM is designed to obtain the channel weights of the two modalities. The difference from FSM is that the two input features are concatenated here, and GAP and global maximum pooling (GMP) are used to obtain global

information and salient information. Finally, two fully connected layers and a softmax function to transform the channel vector nonlinearly obtain the channel weight w_c.

$$w_c = \phi_2(\phi_1(GAP(P^a) + GMP(P^a)) \tag{6}$$

$$P^a = cat(x_i, x_v) \tag{7}$$

where ϕ_1, ϕ_2 denote the fully connected layer. P^a represents the concatenated feature. x_i and x_v represent the output of the third layer.

At the same time, we design a spatial integration branch, which uses the max function to obtain the maximum value of the corresponding channel dimension of each spatial position and uses a 3×3 convolution smoothing feature to get the spatial weight w_s. w_s evaluates the contribution of each spatial location while highlighting candidates and suppressing distractions.

$$w_s = conv_{3\times3}(max(P^a)) \tag{8}$$

After the channel and spatial weights are obtained, they are fixed between 0 and 1 through softmax function σ. Additionally, we multiply them to get a 3D feature matrix with the same size as the original feature, then multiply the 3D feature with P^a. Finally, the element-wise summation of the P^a is carried out to get the optimized feature P. The mathematical expression is given as follows:

$$P = P^a \otimes [\sigma(w_s) \otimes \sigma(w_c)] + P^a \tag{9}$$

4. Experiments

To verify the effectiveness of our proposed method, we conducted many representative experiments. The experimental environment is configured as follows: NVIDIA GeForce RTX3090 GPU server, PyTorch 1.12, and Python 3.8.

4.1. Implementation Details

4.1.1. Network Training

In this work, we train MSIFNet in multiple steps, and stochastic gradient descent (SGD) algorithm is adopted to optimize MSIFNet. During training, we use softmax cross-entropy loss function for binary classification. The weight decay and momentum are set to 0.0005 and 0.9, respectively. The first step is to train the backbone network. The backbone network is initialized by the pre-trained VGG-M model on ImageNet [42], and the specific-modality branch and the full connection layer are trained with 100 epoch iterations. The learning rate of the convolution layer and the full connection layer is 0.0005 and 0.001, respectively. In the second step, we loaded the training parameters of the first step, randomly initialized the parameters of the multi-scale branches and FSM, and conducted the training of 500 epoch iterations. The learning rate was consistent with that of the first step. In the third stage, we load the training parameters of the second step and randomly initialize TIFM and GFM to train 1000 epoch iterations, and the learning rate of TIFM is 0.001, and the others are 0.0005. In each iteration, eight frames of images are randomly obtained from the video sequence. Gaussian samples are performed on these 8 frames to get 256 positive and 768 negative samples, of which 32 positive samples and 96 negative samples are generated for each frame image. It is considered a positive sample when the ratio of overlap rate (IoU) between the sample and the ground truth is [0.7, 1], while a sample in the range of [0, 0.5] is considered a negative sample. Notably, we train MSIFNet with the GTOT [43] dataset and test it on RGBT234 [44] and LasHeR [45] datasets. When testing on GTOT, we randomly selected 50 video sequences on RGBT234 for training.

4.1.2. Online Tracking

Like MDNet [46], we freeze all network parameters except Fc4-Fc6 and randomly initialize a new FC6 branch during tracking. Then, Gaussian sampling is performed on the target bounding box in the first frame, and 500 positive samples and 5000 negative

samples are obtained as training sets for 50 epoch iterations of fine-tuning FC4, FC5, and FC6. The learning rate for Fc4 and Fc5 is 0.0001, and for Fc6 is 0.001. To make the tracking results more accurate, we update the parameters of the three fully connected layers in the short- and long-term and use the bounding box regression technique to solve the target scale change problem. Given the unreliability of subsequent frames, we only collected 1000 samples in the first frame to train the regression factor. For the t frame, the network uses the tracking result of the t-1 frame to sample 256 candidate samples and send them to the network to predict the target state. We select the five candidate samples with the highest score and take their average as the tracking result of the current frame. The tracking is successful when the result score exceeds 0, and the results are adjusted using a bounding box regressor for more accurate positioning. The network repeats until the entire sequence is tracked.

4.2. Result Comparisons

4.2.1. Datasets and Evaluation Metrics

(1) GTOT dataset: the GTOT dataset contains 50 pairs of RGBT video sequences collected in different scenarios and conditions, aligned spatially and temporally, totaling about 15K frames. According to the target state, it is divided into 7 challenge attributes to analyze the performance of the tracker under different conditions.
(2) RGBT234 dataset: it is a large dataset after the RGBT210 [47] dataset, adding 34 video sequences based on RGBT210. It includes 234 pairs of highly aligned RGBT video sequences and 12 challenge attributes for approximately 234K frames. It provides more accurate annotations and considers the challenges posed by various environments.
(3) LasHeR dataset: the LasHeR dataset is a more comprehensive and extensive RGBT dataset containing 1224 pairs of aligned video sequences and 19 attribute annotations. Among them, 245 sequences were selected as the test dataset, and the rest were used for training.

In the above datasets, precision rate (PR) and success rate (SR) are commonly used to evaluate tracker performance. Specifically, PR represents the percentage of frames for which the Euclidean distance between the tracking result and the ground truth is below the set threshold. Since the target of the GTOT is smaller, the threshold is set to 5 pixels. We set the threshold to 20 pixels for RGBT234 and LasHeR datasets. SR measures the percentage of frames where the overlap ratio between the tracking result and the ground truth is greater than the set threshold. Normalized PR (NPR) [48] is also commonly used to evaluate the LasHeR dataset.

4.2.2. Evaluation of GTOT Dataset

Overall Comparison: To verify the effectiveness of MSIFNet on the GTOT dataset, we compared the tracking results with some advanced RGBT tracking algorithms, including M^5L [49], CAT [12], APFNet [38], MANet [11], HDINet [50], DAFNet [35], MACNet [51], SGT [47], ECO [52], MDNet+RGBT, RT-MDNet [53], etc. The results of the PR and SR evaluations are shown in Figure 7. Specifically, MSIFNet gains 10.4% in PR and SR compared to the baseline tracker MDNet + RGBT. Compared to the recent APFNet, MSIFNet is 0.4% higher on SR and only 0.1% lower on PR. Compared to HDINet and CAT trackers, our trackers have a performance improvement of 1.6%/2.3% and 1.5%/2.4% on PR/SR, respectively. In summary, compared with some advanced RGBT algorithms, the effectiveness of our algorithm is proven.

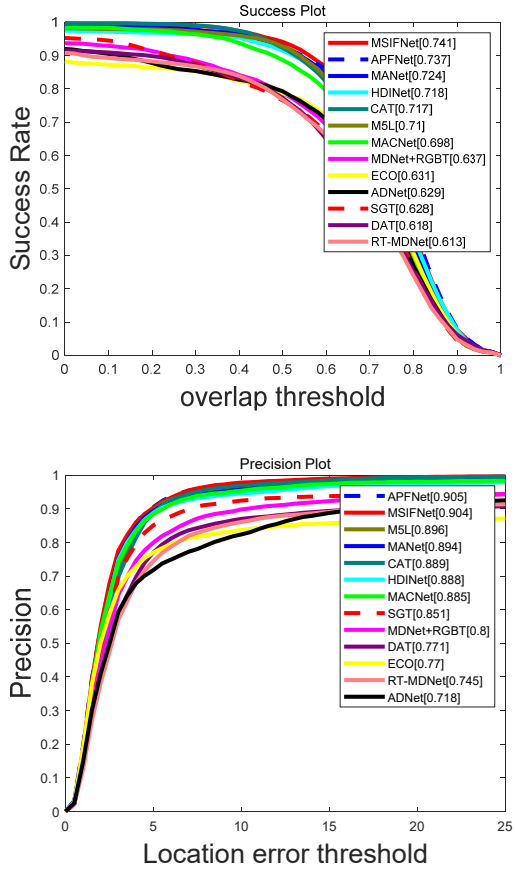

Figure 7. Evaluation curves for PR/SR metrics on GTOT.

Attribute-based performance: To evaluate the specific performance of our algorithm in addressing challenges, we compared the results of MSIFNet with those of other state-of-the-art RGBT trackers on various challenge attributes. The GTOT dataset contains 7 challenge attribute labels, including occlusion (OCC), large scale variation (LSV), a small object (SO), fast motion (FM), deformation (DEF), low illumination (LI), and thermal crossover (TC). The evaluation results are presented in Table 1. The best result is indicated in red, and the second and third are green and blue, respectively. The comparative results show that the overall performance of MSIFNet is optimal, especially in solving challenges such as OCC, LSV, FM, TC, and SO, which fully demonstrates the effectiveness of our method. In particular, MSIFNet ranks first for performance in handling large-scale changes and small object challenges, which proves the effectiveness of our proposed design of multi-scale information mining. MSIFNet is first place in both PR and SR on OCC, demonstrating that the GFM has a solid ability to highlight important information about the target and suppress the background.

Table 1. Attribute-based PR/SR scores (%) on the GTOT dataset.

	SGT	MDNet + RGBT	MANet	CAT	APFNet	MSIFNet
OCC	81.0/56.7	82.9/64.1	88.2/69.6	89.9/69.2	90.3/71.3	90.9/72.2
LSV	84.2/54.7	77.0/57.3	86.9/70.6	85.0/67.9	87.7/71.2	88.0/72.0
FM	79.9/55.9	80.5/59.8	87.9/69.4	83.9/65.4	86.5/68.4	88.2/70.2
LI	88.4/65.1	79.5/64.3	91.4/73.6	89.2/72.3	91.4/74.8	91.2/74.5
TC	84.8/61.5	79.5/60.9	88.9/70.2	89.9/71.0	90.4/71.6	91.4/72.9
SO	91.7/61.8	87.0/66.2	93.2/70.0	94.7/69.9	94.3/71.3	95.9/72.3
DEF	91.9/73.3	81.6/68.8	92.3/75.2	92.5/75.5	94.6/78.0	92.8/77.5
ALL	85.1/62.8	80.0/63.7	89.4/72.4	88.9/71.7	90.5/73.7	90.4/74.1

4.2.3. Evaluation on RGBT234 Dataset

Overall Comparison: As shown in Figure 8, a comprehensive evaluation was performed on the RGBT234 dataset. The algorithm in this paper also achieves advanced performance. Compared with JMMAC [54], the proposed algorithm improves PR by 2.7%, while SR is only 0.3% lower. In addition, compared to CAT, M^{5}L and HDINet, MSIFNet has a performance improvement of 1.3%/0.9%, 2.2%/2.8%, and 3.4%/1.1% on PR/SR, respectively. However, compared to high-performance CMPP [36], MSIFNet has a certain gap. The reason is that the proposed algorithm only uses the information of the current frame to model the target appearance, while CMPP enhances the representation of the current frame with the help of a large amount of historical information.

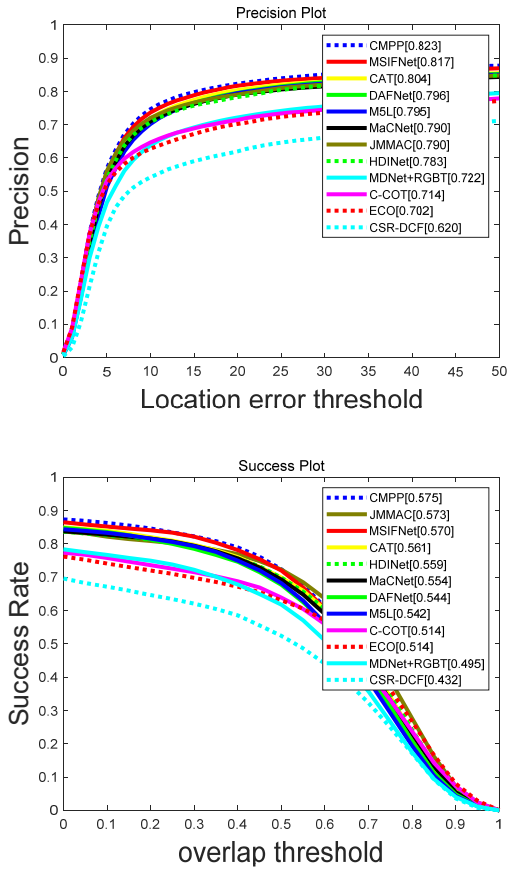

Figure 8. Evaluation curves for PR/SR metrics on RGBT234.

Attribute-based performance: At the same time, we also compared MSIFNet with 5 excellent trackers on 12 challenge attributes on the RGBT234 dataset. Challenging attributes include scale variation (SV), partial occlusion (PO), deformation (DEF), background clutter (BC), low resolution (LR), low illumination (LI), camera movement (CM), fast motion (FM), heavy occlusion (HO), no occlusion (NO), motion blur (MB), and thermal crossover (TC). As can be seen from Table 2, MSIFNet performs well in the challenges of HO, SV, LI, LR, TC, MB, DEF, CM, and FM. Notably, MSIFNet ranked in the top three across all challenges. The PR/SR metric on SV is 82.0%/57.6% compared to state-of-the-art RGBT trackers, proving that MSIFNet can adapt well to scale changes and mine multi-scale clues well. Moreover, the performance of our algorithm on LI is also excellent, which proves that the FSM and TIFM can achieve modality information interaction and fusion well. The FSM and TIFM make the RGB modality have the infrared modality information, which can cope with the challenge of low illumination well.

Table 2. Attribute-based PR/SR scores (%) on the RGBT234 dataset.

	MDNet + RGBT	DAFNet	M^{5}L	CAT	HDINet	MSIFNet
NO	86.2/61.1	90.0/63.6	93.1/64.6	93.2/66.8	88.4/65.1	92.6/66.0
PO	76.1/51.8	85.9/58.8	86.3/58.9	85.1/59.3	84.9/60.4	84.7/59.5
HO	61.9/42.1	68.6/45.9	66.5/45.0	70.0/48.0	67.1/47.3	73.8/50.5
LI	67.0/45.5	81.2/54.2	82.1/54.7	81.0/54.7	77.7/53.2	83.2/55.7
LR	75.9/51.5	81.8/53.8	82.3/53.5	82.0/53.9	80.1/54.5	85.1/56.5
TC	75.6/51.7	81.1/58.3	82.1/56.4	80.3/57.7	77.2/57.5	84.4/59.0
DEF	66.8/47.3	74.1/51.6	73.6/51.1	76.2/54.1	76.2/56.5	74.7/53.5
FM	58.6/36.3	74.0/46.5	72.8/46.5	73.1/47.0	71.7/47.5	74.7/47.5
SV	73.5/50.5	79.1/54.4	79.6/54.2	79.7/56.6	77.5/55.8	82.0/57.6
MB	65.4/46.3	70.8/50.0	73.8/52.8	68.3/49.0	70.8/52.6	75.8/54.4
CM	64.0/45.4	72.3/50.6	75.2/52.9	75.2/52.7	69.7/51.4	77.3/54.9
BC	64.4/43.2	79.1/49.3	75.0/47.7	81.1/51.9	71.1/47.8	80.5/52.3
ALL	72.2/49.5	79.6/54.4	79.5/54.2	80.4/56.1	78.3/55.9	81.7/57.0

4.2.4. Evaluation of LasHeR Dataset

To further demonstrate the robustness of MSIFNet to different datasets, we evaluate the performance of the tracker on LasHeR. The comparison results of NPR and SR indicators are shown in Figure 9. Compared to 11 RGBT algorithms, MSIFNet obtained the best results in NPR and SR. Specifically, MSIFNet outperformed the best-performing MaCNet on NPR/SR, up 0.8% and 0.2%, respectively. In addition, MSIFNet has a 3.3%/3.8% and 2.4%/3.8% improvement in NPR and SR compared to CAT and MANet++ [55]. These results demonstrate the robustness of MSIFNet to different datasets.

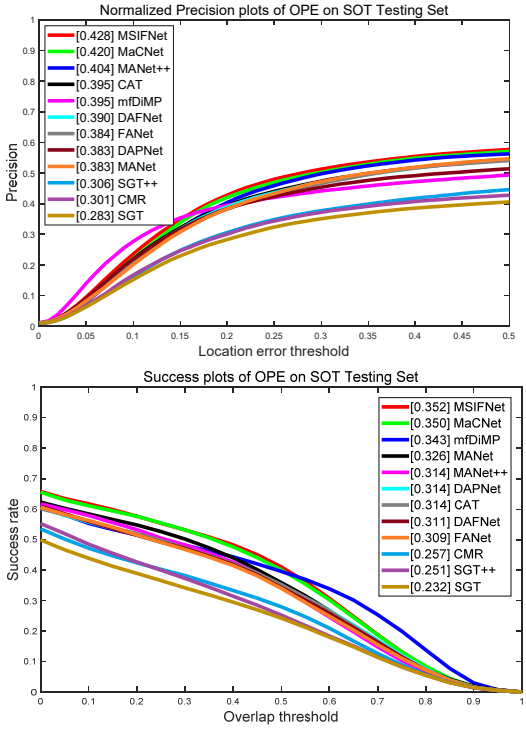

Figure 9. Evaluation curves for NPR/SR metrics on LasHeR.

4.3. Qualitative Analysis

In Figure 10, we perform a qualitative analysis of our algorithm. We compared MSIFNet with four advanced RGBT algorithms on six pairs of video sequences, namely children4, mancross1, diamond, Yellowcar, single3, and Night2. In the case of the first children4 sequence target moving the fast and low resolution, the target bounding box of all other trackers gradually drifts, and the accuracy of MSIFNet is barely affected. In

the mancross1 sequence and diamond sequence, there are challenging attributes such as background clutter, scale variation, and heavy occlusion. Other algorithms keep the target in the lost state after heavy occlusion, but our method can track the object in a good state after occlusion. Meanwhile, in the Yellowcar sequence, our tracker can still maintain almost the same size as the ground truth box as the target scale changes. Our tracker can also robustly track the target for the single3 and Night2 sequences with a small object, fast motion, low illumination, background clutter, and partial occlusion attributes. It is clear that MSIFNet can locate the target accurately during the tracking process and is capable enough to cope with various challenges.

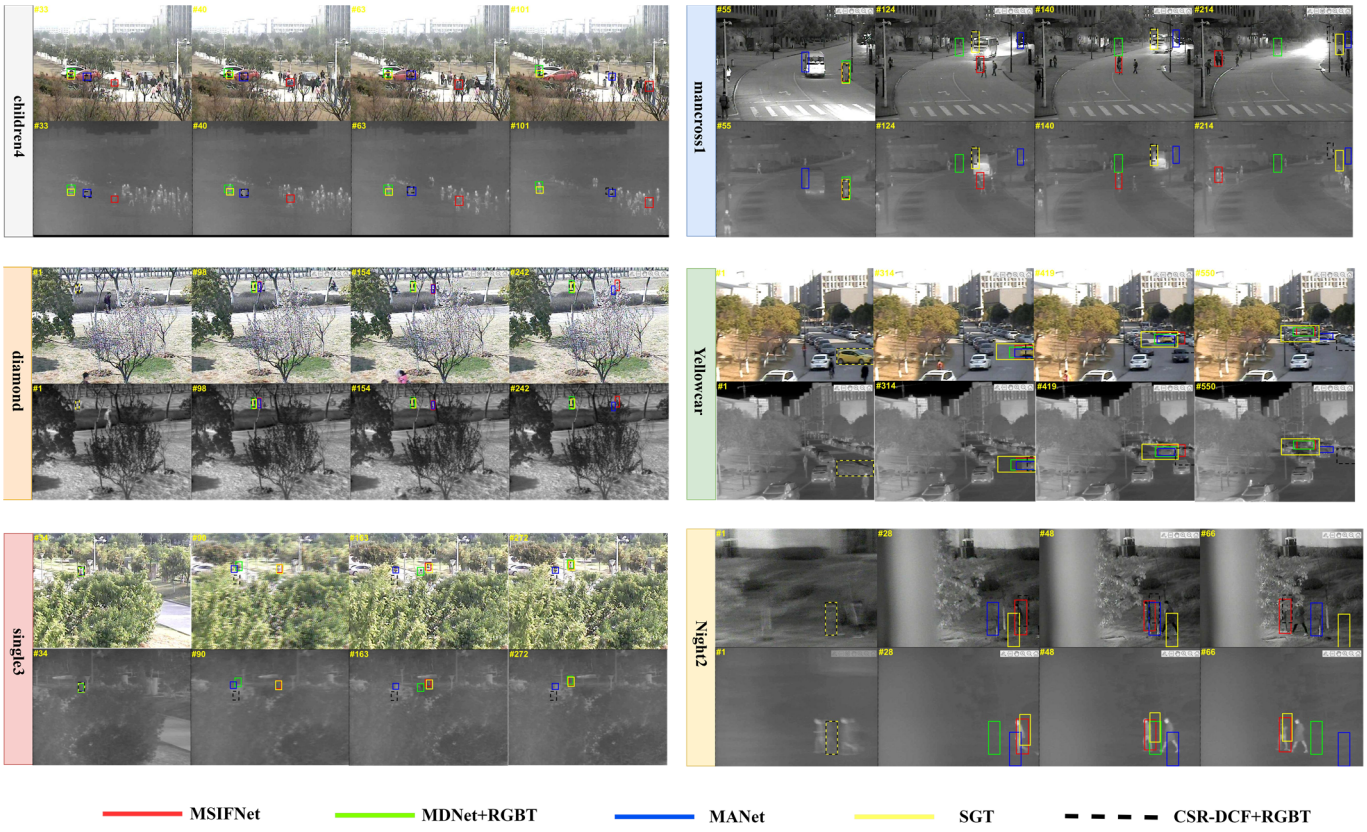

Figure 10. Qualitative comparison with four excellent RGBT trackers.

4.4. Ablation Study

To verify the validity of the proposed modules, ablation experiments were performed on the RGBT234 dataset. We compared the tracking performance of the following methods: (1) MSIFNet-FSM, which removes the FSM and performs simple addition operations for multi-scale features; (2) MSIFNet-TIFM, which removes the TIFM and adds the output of FSM directly to the backbone features; (3) MSIFNet-GFM, which removes the GFM and directly concatenates the output features of the third layer; and (4) MSIFNet-Gate, which removes the Gated network and inputs the TIFM for interactive fusion without gating. As seen from Table 3, MSIFNet-FSM has the lowest performance, mainly because a large amount of redundant information is inevitably introduced when aggregating data from multiple branches, which proves the necessity of introducing aggregation modules into multi-scale branches. The performance of MSIFNet is significantly better than that of MSIFNet-TIFM, which verifies that the proposed TIFM can effectively enhance semantic information by learning long-distance dependencies. The accuracy of MSIFNet-GFM and MSIFNet-Gate is somewhat reduced in MSIFNet, which also demonstrates the effectiveness of the GFM and Gate network.

Table 3. The PR/SR scores of our method were compared with variants on the RGBT234 datasets to verify the validity of the proposed modules.

		MSIFNet-FSM	MSIFNet-TIFM	MSIFNet-GFM	MSIFNet-Gate	MSIFNet
RGBT234	PR	0.799	0.806	0.813	0.811	**0.817**
	SR	0.557	0.561	0.564	0.565	**0.570**

We compared the performance of the following variants to verify the effectiveness of the multi-scale features extracted from each layer. (1) MSIFNet-L1 means deleting the 5×5 and 3×3 branches of the first layer and directly fusing the backbone features while leaving the other layers unchanged. (2) MSIFNet-L2 means that the 3×3 and 1×1 branches of the second layer are deleted, while others remain unchanged. (3) MSIFNet-L3 means to delete the convolution branch of 1×1 in the third layer. The results of the comparison are shown in Table 4. Experimental results show that the performance of MSIFNet is better than that of MSIFNet-L1, MSIFNet-L2, and MSIFNet-L3, which proves the effectiveness of multi-scale features in each layer.

Table 4. The PR/SR scores of our method were compared with variants on the RGBT234 datasets to verify the validity of the multi-scale features extracted from each layer.

		MSIFNet-L1	MSIFNet-L2	MSIFNet-L3	MSIFNet
RGBT234	PR	0.801	0.806	0.814	**0.817**
	SR	0.563	0.563	0.568	**0.570**

5. Conclusions

In this paper, we propose a multi-scale feature interaction fusion network (MSIFNet), which can mine the multi-scale information of RGB and infrared images to better identify targets of different sizes and further improve tracking performance. In particular, we use a feature selection module to adaptively select features from multiple branches, and the Transformer interactive fusion module is used to mine complementary information between features and enhance semantic representation. Moreover, a global feature fusion module is designed to adjust the global information in spatial and channel dimensions, respectively. Numerous experiments on publicly available GTOT, RGBT234, and LasHeR datasets have shown that our MSIFNet has advanced performance in handling various challenges. In the future, we will further explore multi-modalities fusion mechanisms to achieve more robust feature representation.

Author Contributions: Conceptualization, X.X. (Xingzhong Xiong); methodology, X.X. (Xianbing Xiao); investigation, X.X. (Xianbing Xiao); writing—original draft preparation, X.X. (Xianbing Xiao); writing—review and editing, X.X. (Xingzhong Xiong), F.M. and Z.C.; supervision, X.X. (Xingzhong Xiong), F.M. and Z.C.; project administration, X.X. (Xingzhong Xiong) and F.M.; funding acquisition, X.X. (Xingzhong Xiong) and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Department of Sichuan Province (2023NSFSC1987) and in part by the Postgraduate Innovation Fund Project of Sichuan University of Science and Engineering under grant Y2022147 and the Artificial Intelligence Key Laboratory of Sichuan Province (grant number 2019RZJ04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Public datasets.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Shen, Y.; Liu, Z.; Zhang, G. PAC interaction inspection using real-time contact point tracking. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 4051–4064. [CrossRef]
2. Mehmood, K.; Jalil, A.; Ali, A.; Khan, B.; Murad, M.; Khan, W.U.; He, Y. Context-aware and occlusion handling mechanism for online visual object tracking. *Electronics* **2020**, *10*, 43. [CrossRef]
3. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]
4. Schnelle, S.R.; Chan, A.L. Enhanced target tracking through infrared-visible image fusion. In Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 5–8 July 2011; IEEE: New York, NY, USA, 2011; pp. 1–8.
5. Chan, A.L.; Schnelle, S.R.J.O.E. Fusing concurrent visible and infrared videos for improved tracking performance. *Opt. Eng.* **2013**, *52*, 017004. [CrossRef]
6. 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]
7. Xia, W.; Zhou, D.; Cao, J.; Liu, Y.; Hou, R. CIRNet: An improved RGBT tracking via cross-modality interaction and re-identification. *Neurocomputing* **2022**, *493*, 327–339. [CrossRef]
8. Lu, A.; Qian, C.; Li, C.; Tang, J.; Wang, L. Duality-gated mutual condition network for RGBT tracking. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: New York, NY, USA, 2022.
9. He, F.; Chen, M.; Chen, X.; Han, J.; Bai, L. SiamDL: Siamese Dual-Level Fusion Attention Network for RGBT Tracking. *SSRN*, 2022; *submitted*. [CrossRef]
10. Wang, Y.; Wei, X.; Tang, X.; Wu, J.; Fang, J. Response map evaluation for RGBT tracking. *Neural Comput. Appl.* **2022**, *34*, 5757–5769. [CrossRef]
11. Long Li, C.; Lu, A.; Hua Zheng, A.; Tu, Z.; Tang, J. Multi-adapter RGBT trackin. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
12. Li, C.; Liu, L.; Lu, A.; Ji, Q.; Tang, J. Challenge-aware RGBT tracking. In *Part XXII 16, Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 222–237.
13. 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.
14. Xu, Q.; Mei, Y.; Liu, J.; Li, C. Multimodal cross-layer bilinear pooling for RGBT tracking. *IEEE Trans. Multimedia* **2021**, *24*, 567–580. [CrossRef]
15. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
16. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Lasvegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
17. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
18. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef]
19. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
20. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *Part II 14, Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10, 15–16 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
21. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
22. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 4282–4291.
23. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12549–12556.
24. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677.
25. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6269–6277.
26. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669.
27. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1571–1580.

28. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Van Gool, L. Transforming model prediction for tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8731–8740.
29. 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; IEEE: New York, NY, USA, 2011; pp. 1–8.
30. Li, L.; Li, C.; Tu, Z.; Tang, J. A fusion approach to grayscale-thermal tracking with cross-modal sparse representation. In *Proceedings of the Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, 8–10 April 2018*; Revised Selected Papers 13; Springer: Berlin/Heidelberg, Germany, 2018; pp. 494–505.
31. Lan, X.; Ye, M.; Zhang, S.; Zhou, H.; Yuen, P.C. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognit. Lett.* **2020**, *130*, 12–20. [CrossRef]
32. Li, C.; Zhu, C.; Huang, Y.; Tang, J.; Wang, L. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 808–823.
33. Shen, L.; Wang, X.; Liu, L.; Hou, B.; Jian, Y.; Tang, J.; Luo, B. RGBT tracking based on cooperative low-rank graph model. *Neurocomputing* **2022**, *492*, 370–381. [CrossRef]
34. Xu, N.; Xiao, G.; Zhang, X.; Bavirisetti, D.P. Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences. In Proceedings of the 4th International Conference on Virtual Reality, Hong Kong, China, 24–26 February 2018; pp. 44–49.
35. 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 IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
36. Wang, C.; Xu, C.; Cui, Z.; Zhou, L.; Zhang, T.; Zhang, X.; Yang, J. Cross-modal pattern-propagation for RGB-T tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7064–7073.
37. 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]
38. Xiao, Y.; Yang, M.; Li, C.; Liu, L.; Tang, J. Attribute-based progressive fusion network for rgbt tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2022; pp. 2831–2838.
39. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
40. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In Proceedings of the British Machine Vision Conference 2014, Nottinghamshire, UK, 1–5 September 2014.
41. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
43. 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]
44. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
45. 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.* **2021**, *31*, 392–404. [CrossRef] [PubMed]
46. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4293–4302.
47. 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, New York, NY, USA, 23–27 October 2017; pp. 1856–1864.
48. Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; Ghanem, B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 300–317.
49. Tu, Z.; Lin, C.; Zhao, W.; Li, C.; Tang, J. M 5 l: Multi-modal multi-margin metric learning for RGBT tracking. *IEEE Trans. Image Process.* **2021**, *31*, 85–98. [CrossRef]
50. 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]
51. Zhang, H.; Zhang, L.; Zhuo, L.; Zhang, J. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors* **2020**, *20*, 393. [CrossRef]
52. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
53. 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.

54. Zhang, P.; Zhao, J.; Bo, C.; Wang, D.; Lu, H.; Yang, X. Jointly modeling motion and appearance cues for robust RGB-T tracking. *IEEE Trans. Image Process.* **2021**, *30*, 3335–3347. [CrossRef] [PubMed]
55. Lu, A.; Li, C.; Yan, Y.; Tang, J.; Luo, B. RGBT tracking via multi-adapter network with hierarchical divergence loss. *IEEE Trans. Image Process.* **2021**, *30*, 5613–5625. [CrossRef] [PubMed]