mathematics

MDPI

Article

Attention and Pixel Matching in RGB-T Object Tracking

Da Li, Yao Zhang, Min Chen * and Haoxiang Chai

School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; frankli26@whut.edu.cn (D.L.); zyao@whut.edu.cn (Y.Z.); hx453794261@whut.edu.cn (H.C.)
* Correspondence: minch@whut.edu.cn

Abstract: Visual object tracking using visible light images and thermal infrared images, named RGB-T tracking, has recently attracted increasing attention in the tracking community. Deep neural network-based methods becoming the most popular RGB-T trackers, still have to balance the robustness and the speed of calculation. A novel tracker with Siamese architecture is proposed to obtain the accurate object location and meet the real-time requirements. Firstly, a multi-modal weight penalty module is designed to assign different weights to the RGB and thermal infrared features. Secondly, a new pixel matching module is proposed to calculate the similarity between each pixel on the search and the template features, which can avoid bringing excessive background information versus the regular cross-correlation operation. Finally, an improved anchor-free bounding box prediction network is put forward to further reduce the interference of the background information. The experimental results on the standard RGB-T tracking benchmark datasets show that the proposed method achieves better precision and success rate with a speed of over 34 frames per second which satisfies the real-time tracking.

Keywords: RGB-T tracking; weight penalty; pixel matching; anchor-free

MSC: 68T01; 68T07; 68T45

Citation: Li, D.; Zhang, Y.; Chen, M.; Chai, H. Attention and Pixel Matching in RGB-T Object Tracking. *Mathematics* **2023**, *11*, 1646. https://doi.org/10.3390/math11071646

Academic Editors: Fan Zhang, Songhe Feng, Yongsheng Zhou and Junlin Hu

Received: 9 March 2023
Revised: 27 March 2023
Accepted: 27 March 2023
Published: 29 March 2023

1. Introduction

Visual object tracking is a significant research direction in the field of computer vision, with widespread applications in domains such as video surveillance, unmanned aerial vehicle (UAV) navigation, and other related fields [1]. The objective of visual object tracking is to estimate the position and dimensions of an object bounding box throughout a complete video sequence based on the ground truth box of the initial frame. Most tracking methodologies rely on RGB images, which employ the extracted features of RGB images to estimate and predict the position of objects. Nonetheless, the outcomes obtained solely from the information provided by visible light images are not optimal in challenging scenarios with limited lighting conditions, such as during night-time, rain, and fog.

RGB-T object tracking is a methodology that incorporates information from both RGB images and thermal infrared (TIR) images to predict the location of an object [2]. RGB and TIR images possess complementary attributes. RGB images are susceptible to illumination changes while providing more detailed information, and TIR images are unaffected by illumination while lacking texture and detail information. Consequently, RGB-T object tracking that leverages the complementary features of RGB and TIR images exhibits superior performance.

Currently, MDNet-based [3] RGB-T trackers have achieved good tracking accuracy [4–7], but the tracking speed of these trackers cannot meet the requirements of real time; Siamese-based trackers are faster and can meet real-time requirements [8–10], but there is a certain gap in performance compared with some advanced trackers. Hence, most of the current RGB-T trackers are difficult to simultaneously satisfy the requirements of robustness and speed. We

designed an RGB-T tracker based on the Siamese network, which can achieve good accuracy while meeting the real-time speed requirements.

Our contributions can be summarized as follows:

1. A multi-modal weight penalty module is proposed to fully use the advantages of the two modal features and deal with various complex illumination challenges.
2. A pixel-matching module and an improved anchor-free position prediction network are proposed to suppress the influence of cluttered background on the localized object and locate the object accurately and quickly for tracking.
3. A new end-to-end RGB-T tracker based on Siamese-net is proposed, which can satisfy the robustness and real-time tracking. The experimental results on two standard datasets show our new tracker is effective.

The rest of this paper is arranged as follows. Section 2 reviews some related works on RGB-T object tracking and briefly introduces some existing RGB-T trackers. In Section 3, our new tracker is described in detail from three aspects. Section 4 the related experiments and results are present for the proposed tracker. Section 5 discusses the advantages and disadvantages of our tracker. In Section 6, the conclusion and future work are given.

2. Related Works

RGB-T object tracking algorithms can be divided into traditional algorithms, algorithms based on correlation filtering, and algorithms based on deep learning. Traditional algorithms mainly employ hand-crafted features such as Histogram of Oriented Gradient (HOG), Scale-Invariant Feature Transform (SIFT), coupled with motion estimation algorithms such as Kalman filter [11] and particle filter [12], to achieve tracking.

Correlation filter-based algorithms obtain the output response by correlating the filter template with the object candidate region features and determining the object position according to the response peak. Zhai et al. [13] proposed an RGB-T tracking algorithm based on a cross-pattern correlation filter, in which correlation filtering was applied to each modality. A low-rank constraint was introduced to jointly learn the filter for modality collaboration. Yun et al. [14] proposed a discriminative fusion correlation learning model, which obtained the fusion learning filter through early estimation and late fusion, to improve the tracking performance of a discriminative correlation filter. Xiong et al. [15] proposed an RGB-T dual-modal weighted correlation filter tracking algorithm in which a weight map was jointly solved by dual-modal information, and the weight map guided the solution of the correlation filter for inferring the object occlusion. Due to the limited representational capacity of hand-crafted features, the accuracy and robustness of these two tracking algorithms are affected.

Deep learning algorithms based on data-driven deep neural networks build trackers with powerful feature representation capabilities, significantly improving accuracy and robustness compared to the previous two algorithms. Xu et al. [16] proposed an RGB-T pixel-level fusion tracking algorithm based on convolutional neural networks, taking thermal infrared images as the fourth channel of visible light images. Li et al. [17] proposed a multi-adapter convolutional network, which extracts shared and specific information of RGB-T bimodally through different adapters. ZHU et al. [18] designed a feature aggregation and pruning module for the RGB-T tracking network. The aggregation module provides rich RGB-T feature representation for the target object, and the pruning module removes noise or redundant features from the aggregated RGB-T features. Lu et al. [19] proposed a dual-selection conditional network to fully utilize bimodal discriminative information and suppress data noise and designed a flow-based resampling strategy to cope with sudden camera movement. Due to the additional modality, additional computations are required. The above algorithms mostly adopt the idea of generating candidate regions, which requires multiple forward propagations of the network, thus affecting the tracking speed.

In recent years, Siamese networks have achieved high accuracy and speed in visible light object tracking, such as SiamFC [20], SiamRPN++ [21], and SiamBAN [22], and the

same type of method has also been applied to RGB-T object tracking. ZHANG et al. [23] proposed a deep learning tracking method based on pixel-level fusion, which first fused visible light and thermal infrared images, and then inputted Siamese networks for tracking. Zhang et al. [24] extracted features of visible light and thermal infrared images separately using two Siamese networks and then fused multi-layer features of the bimodal system and used multi-layer fusion features for tracking.

3. Method

In this section, we first describe the Siamese network architecture of the RGB-T tracker in detail and introduce the improved anchor-free position prediction network. Then, we introduce the two main modules of our tracker, including the multi-modal weight penalty module and the pixel-matching module.

3.1. Siamese Network Architecture

The overall network framework of our RGB-T Siamese tracker is shown in Figure 1. The Siamese network comprises a template branch and a search area branch. Each branch is divided into two branches: the RGB images branch and the thermal infrared images branch, which extract the features of these two images, respectively. The two branches have the same structure and share parameters. The resnet50 combined with FPN [25] is employed as the backbone for feature extraction to obtain the features extracted from the second, third, and fourth convolutional layers. Firstly, the multi-modal feature penalty module weights the RGB and TIR modality features. Then, the template and search features of the two modalities are sent to the pixel-matching module for the matching operation, and then the response maps of the two modalities are fused. The fused maps are directed into a regression and classification subnetwork similar to SiamCAR [26]. Diverging from SiamCAR, distinct strategies are employed for positive sample determination, and improvements are made to the regression branch.

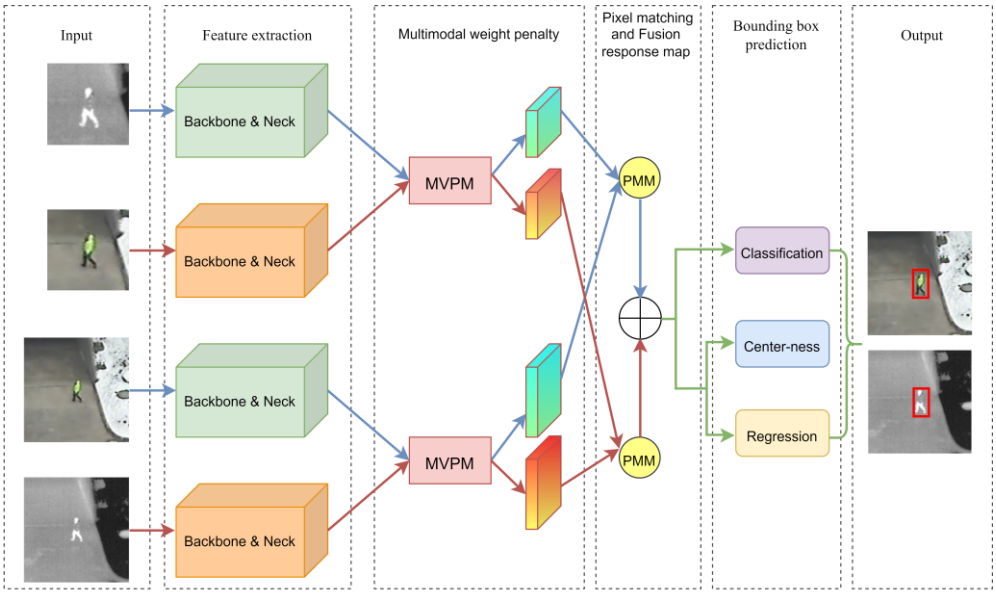

Figure 1. The overall architectural diagram of the proposed algorithm. The overall network consists of four main components: a double Siamese network for feature extraction, an MWPM for enhancing multi-modal features, a PMM for generating fusion response maps, and an anchor-free position prediction network for generating bounding boxes.

When the location (x, y) falls into the center region of any ground truth box, it is considered as a positive sample, as shown in the red part in Figure 2. The part inside the marked box near the edge often belongs to the background part, and this part of the points

that belong to the negative sample is incorrectly classified as the positive sample. This will lead to classification errors in the classification branch, causing trouble in learning the model. In addition, for this anchor-free method, the convolutional features in the central part are richer, and the object edge can be better predicted only by the features in the central part. Therefore, we choose the center region of the box centered at (c_x, c_y) to be defined as the subframe $(c_x - rs, c_y - rs, c_x + rs, c_y + rs)$, where s is the stride of the FPN layer currently in place and r is the hyperparameter of 1.5, as shown in the yellow part of Figure 2. In this way, most anchor points close to the edge inside the label box and falling on the background are labeled as negative samples so that our model converges faster and the final performance is better.

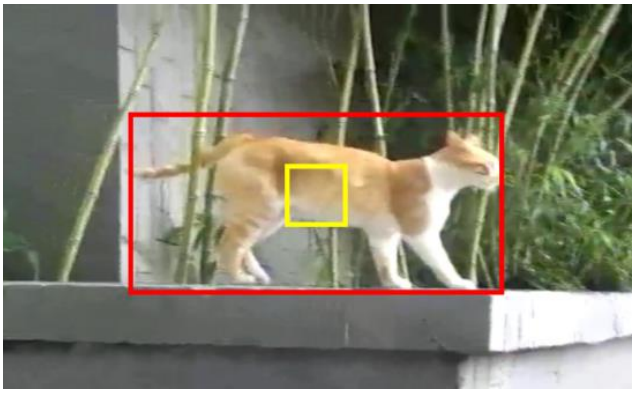

Figure 2. Schematic diagram of the selected positive samples. The red part is the originally defined positive sample area, and the yellow part is our improved positive sample area.

The regression branch of the network produces a regression feature map A_{reg}, which has dimensions of $25 \times 25 \times 4$. Each position (i, j) in A_{reg} can be mapped back to the corresponding location in the search area (x, y). The regression objective of A_{reg} at (i, j) is a four-dimensional vector $t_{(i,j)} = (l, t, r, b)$, which can be calculated as:

$$l = \frac{x - x_0}{s}, r = \frac{x_1 - x}{s}, t = \frac{y - y_0}{s}, b = \frac{y_1 - y}{s} \tag{1}$$

l, t, r, b are the distances from position (x, y) to the four sides of the object bounding box. (x_0, y_0) and (x_1, y_1) denote the upper left and lower right corners of the ground truth bounding box. To better adapt to the size of the FPN, the total steps before the feature mapping are increased.

Using the four-dimensional vector $t_{(i,j)}$, the Generalized Intersection over Union (GIOU) metric can be computed to measure the similarity between the predicted and ground truth bounding boxes. The GIOU comprehensively assesses the spatial overlap between the predicted and true bounding boxes, considering their size, location, and shape. Then, the reg loss can be obtained as:

$$L_{reg} = \frac{1}{\sum I\left(t_{(i,j)}\right)} \sum_{i,j} I\left(t_{(i,j)}\right) L_{GIOU}\left(A_{reg}(i,j), t_{(x,y)}\right) \tag{2}$$

The L_{GIOU} denotes the GIOU loss function, as defined in [27], whereas the function $\mathrm{I}(\cdot)$ corresponds to the indicator function introduced in [26].

The final score $S_{x,y}$ at tracking is defined as the square root of the product of the predicted centrality $Cen_{x,y}$ and the corresponding classification score $Cls_{x,y}$, as in Equation (3).

$$S_{x,y} = \sqrt{Cls_{x,y} \times Cen_{x,y}} \tag{3}$$

3.2. Multi-Modal Weight Penalty Module (MWPM)

The features of RGB images and thermal infrared images extracted from the backbone have different effects on object tracking. Previous research on the attention module for RGB object tracking focused on the importance of the channel of each feature, assigning higher weights to relatively larger contribution feature channels and lower weights to smaller contribution features. These methods successfully utilize mutual information from different dimensional features. However, they lack consideration of the contributing factors of weights, which can further suppress unimportant channels or pixels. We use the contributing factors of the weights to represent the contribution of each modal feature. Batch normalization scale factor is used, which uses standard deviations to represent the importance of the weights. To comprehensively consider the contribution of all the features in RGB and TIR modalities towards object representation, we design a multi-modal weight penalty module, which evaluates the features of the two modalities as a whole and then assigns the corresponding weights to the deep features.

The MWPM utilizes the variance of training model weights to highlight salient features. Compared with previous attention mechanisms, no additional calculations and parameters are required, such as full connection and convolution. The scaling factor in batch normalization (*BN*) [28] is directly used to calculate the attention weight, and the non-significant features are further suppressed by adding regularization terms. For the channel attention submodule, the scaling factor reflects the magnitude of the change in each channel and indicates the channel's importance, as shown in Equation (4).

$$BN(x_i) = \gamma \hat{x}_i + \beta \tag{4}$$

where $\hat{x}_i$ is the normalized eigenvalue, and γ and β are the learnable reconstruction parameters that allow our network to learn to recover the feature distribution of the original network. The scaling factor is the variance in *BN*. The larger the variance indicates that the more dramatic the channel changes, the richer the information contained in the channel will be and the greater the importance will be, while the channel information with little change is singular and less important.

$$C_{out} = \frac{\gamma_{c_i}}{\sum_{j=0}^{c} \gamma_{ci}} (BN(C_{in})) \tag{5}$$

$$S_{out} = \frac{\gamma_{s_i}}{\sum_{j=0}^{N} \gamma_{si}} (BN(S_{in})) \tag{6}$$

The channel attention is shown in Equation (5), C_{in} denotes the input features, C_{out} denotes the final obtained output features, and γ_{c_i} is the scaling factor for each channel. If the same normalization method is used for each pixel in space, the spatial attention Equation (6) can be obtained.

As shown in Figure 3, in the channel attention module, RGB and TIR features are first concatenated along the channel dimension to obtain a joint feature representation of the template and search features:

$$U_F = cat(F_{RGB}, F_T) \tag{7}$$

Then, the feature weight penalty is applied to U_F and the output obtained is:

$$(F_v, F_i) = Split(\delta(\frac{\gamma_{c_i}}{\sum_{j=0}^{c} \gamma_{ci}} (BN(U_F)))) \tag{8}$$

where $\delta(\cdot)$ denotes sigmoid activation and $Split$ is an operation that splits features along the channel dimension. For the integration method for two submodules, we adopt a method in which the channel attention is in front and the spatial attention is behind. Firstly, the channel attention module is used to reduce the weight of the less significant feature channels, and then the spatial attention module is used to suppress the background noise.

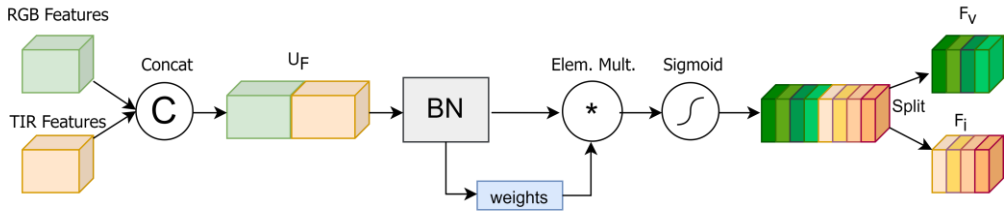

Figure 3. The architecture of MWPM. First, connect the depth features of RGB mode and thermal infrared mode and then assign weights to all channels.

3.3. Pixel Matching Module (PMM)

The RGB object tracking algorithm based on the Siamese network has good performance in accuracy and speed, but when matching template features and search features, their commonly used cross-correlation operations will bring a lot of background information in the deep network, resulting in inaccurate matching. One observation is that in RGB-T tracking tasks, the objects to be tracked are mostly small, especially the GTOT [29] and RGBT234 [30] datasets, and a relatively large template is taken to obtain robust features. Therefore, when a template image of 127 × 127 is inputted, the proportion of objects in the template is very small. With the increase in network depth, especially in deep networks such as ResNet50 [31], even the feature points in the final output correspond to a large receptive field of the input. The size of template features is large, and the corresponding true matching region is much larger than the ideal matching region. Therefore, a large amount of background information will be introduced and overwhelm the features of the object, making it difficult to distinguish the object from similar objects in the background.

To solve the above problems, we use the pixel matching module to calculate the similarity between each pixel on the search and template features. The template feature is transformed into a 1 × 1 kernel, so that the matching area is only of size 1 × 1 and the background information is greatly reduced. The spatial kernel focuses on information from each region of the template, while the channel kernel pays more attention to the overall information of the template. The template feature is decomposed into space and channel kernels with a size of 1 × 1, which reduces the matching area while suppressing background interference and accurately collects the response points of the target area.

$$P_1(i,j)[m] = X_f(i,j) \cdot Z_{fs}^m, \quad m = 1, 2 \ldots, n \tag{9}$$

$$P_2(i,j)[n] = P_1(i,j) \cdot Z_{fc}^m, \quad n = 1, 2 \ldots, c \tag{10}$$

As shown in Figure 4, the template feature Z with the dimension of Hz × Wz × C, which is split into H_Z × W_Z 1 × 1 × C kernels and C 1 × 1 × n (n = w_z × h_z) kernels from the spatial and channel aspects, respectively. Then, search frame feature X is passed through these two filters. Firstly, the similarity between each pixel of the search feature and the spatial dimension template feature is calculated, as in Equation (9). Then, the similarity with the channel kernel can be calculated, as in Equation (10). Finally, the output feature P_2 is obtained. This operation does not require convolution, just matrix multiplication, which can improve the speed of our tracker.

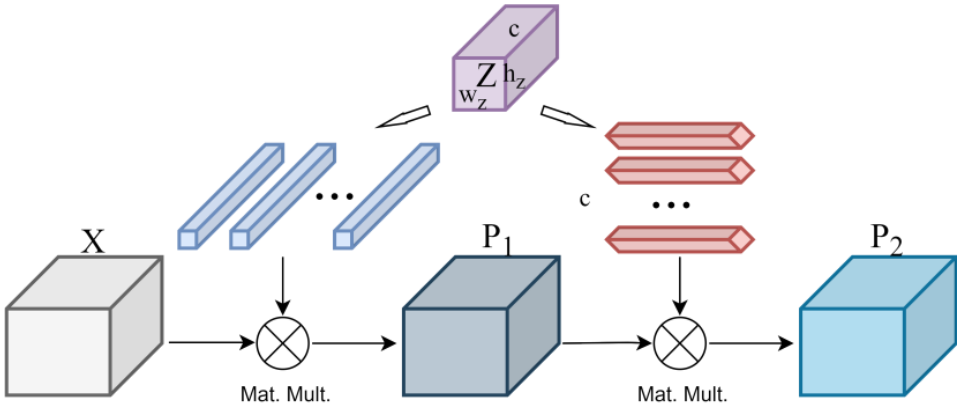

Figure 4. The process of pixel matching. First, calculate the similarity between each pixel point of the search feature and the spatial kernel, and then calculate the similarity between it and the channel kernel.

4. Results

Our network was implemented in python using PyTorch and trained and tested on RTX3060ti based on three publicly benchmark datasets: the GTOT dataset, RGBT234, and LasHeR [32]. We used LasHeR as the training dataset to train our model and then tested it on GTOT and RGBT234, respectively. To evaluate the algorithms, this paper used two common evaluation indicators, precision rate (PR) and success rate (SR). PR represents the percentage of frames whose distance from the center of the prediction box to the center of the ground truth box in the video sequence is less than a predetermined threshold. The threshold for the GTOT dataset is 5 pixels, and that for the RGBT234 dataset is 20 pixels. SR represents the percentage of frames whose overlap rate between the predicted output position box and the ground truth box is greater than the threshold. When the overlap threshold is greater than 0.7, it indicates successful tracking. We compared the performance of the proposed tracker with the advanced RGB-T tracker and RGB tracker, evaluated our main components, and analyzed their effectiveness.

4.1. Results on GTOT

We tested our tracker on the GTOT dataset and compared it with seven advanced trackers, including MANet [17], DAPNet [18], MACNet [33], ECO [34], DAT [35], SGT [36], and SiamDW [37].

Figure 5 shows the graphs of our tracker compared with other trackers on the dataset GTOT. The curves with different colors and lines in the figure represent different trackers. It can be clearly seen that our tracker outperforms the other seven trackers. Specifically, the PR of the tracker proposed in this paper is 1.6% higher than MANet, and the SR differs from MANet by 0.2%; the PR and SR are 2.5% and 1.7% higher than MACNet, respectively. This proves that our proposed tracker can achieve robust tracking. In addition, our proposed tracker achieves a very efficient operating speed. Compared with the better-performing trackers, MANet, DAPNet, MACNet, etc., our tracker achieves the leading efficiency with a real-time operation speed of 34 FPS, which proves that our proposed tracker can achieve the real-time standard for object tracking.

The GTOT dataset contains seven different attributes: occlusion (OCC), large-scale variation (LSV), fast motion (FM), low illumination (LI), thermal crossover (TC), small object (SO), and deformation (DEF). The attribute-based comparison shows the capability of our proposed tracker to handle these challenging situations. As shown in Table 1, our tracker achieves a definite lead in performance under the challenges of large-scale variation and low illumination while obtaining the best performance overall. In addition, our tracker also shows high performance under four attributes: fast motion, thermal crossover, deformation, and small objects. This shows that our tracker can continuously adapt to the changes of the

object during tracking using the pixel matching module and improved full convolutional anchor-free tracking head to reduce the negative impact of background information, thus improving the robustness of object tracking. At the same time, it shows that the introduction of a multi-modal weight penalty module can fully exploit the information of the two modes of RGB-T and make use of the interaction of dual-modal information to better cope with the problem of motion blur caused by fast motion and poor single-modal imaging caused by environmental factors.

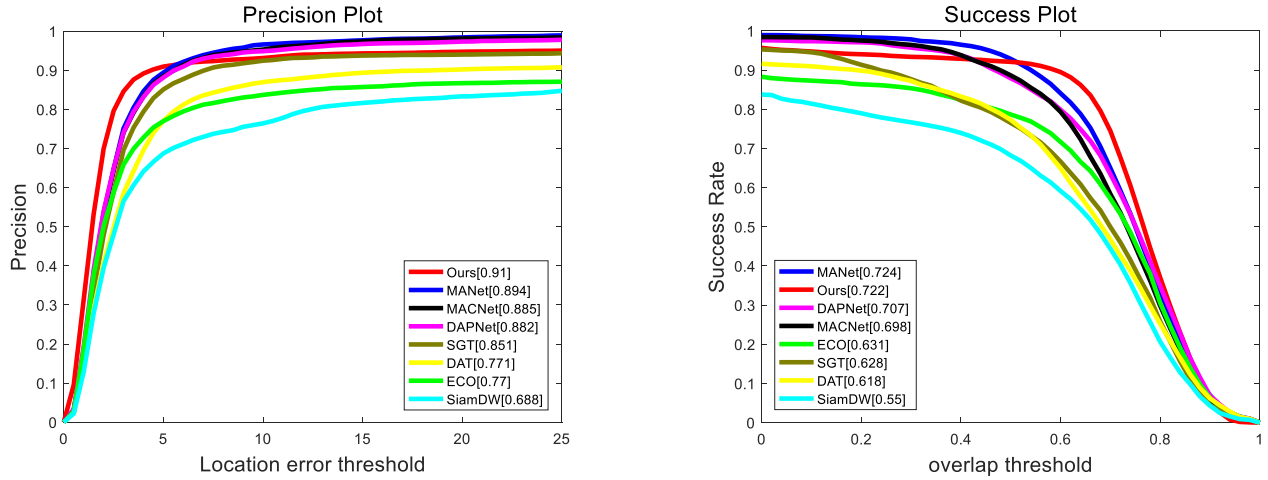

Figure 5. Precision rate (PR) and success rate (SR) curves of different tracking results on GTOT, where the representative PR and SR scores are presented in the legend.

Table 1. Attribute-based accuracy and success rate (PR/SR) is obtained by using different trackers on the GTOT dataset. Red, green, and blue numbers represent the best one, the second-best one, and the third-best one, respectively.

Attributes	**OOC**	**LSV**	**FM**	**LI**	**TC**	**SO**	**DEF**	**ALL**
SiamDW	63.4/49.2	72.0/55.7	63.2/48.4	68.8/55.1	68.4/53.6	73.2/53.4	69.8/55.9	68.8/55.0
ECO	77.5/62.2	85.6/70.5	77.9/64.5	75.2/61.7	81.9/65.3	90.7/69.1	75.2/59.8	77.0/63.1
DAT	77.2/59.2	78.6/62.4	82.0/61.5	76.0/60.9	80.9/62.6	88.6/64.4	76.9/63.3	77.1/61.8
SGT	81.0/56.7	84.2/54.7	79.9/55.9	88.4/65.1	84.8/61.5	91.7/61.8	91.9/73.3	85.1/62.8
DAPNet	87.3/67.4	84.7/64.8	82.3/61.9	90.0/72.2	89.3/69.0	93.7/69.2	91.9/77.1	88.2/70.7
MACNet	86.7/67.6	84.2/66.2	84.4/63.3	90.1/70.7	89.3/68.6	93.2/67.8	92.2/74.4	88.5/69.8
MANet	88.2/69.6	86.9/70.6	87.9/69.4	91.4/73.6	88.9/70.2	93.2/70.0	92.3/75.2	89.4/72.4
Ours	84.8/67.6	93.2/74.5	87.0/70.0	94.8/73.9	88.9/70.4	93.2/72.0	94.6/73.2	91.0/72.2

4.2. Results on RGBT234

We tested our tracker on the RGBT234 dataset and compared it with eight advanced trackers, including DAPNet, MDNet, SGT, DAT, ECO, C-COT [38], SiamDW + RGBT, and SOWP [39].

Figure 6 shows that the experimental results of our tracker on the RGBT234 dataset are PR = 77.5% and SR = 54.8%, which achieves excellent performance compared with other algorithms. Specifically, the PR of the tracker proposed in this paper is 0.9% higher than that of DAPNet, which ranks second in the figure, and the SR is 1.1% higher than that of DAPNet. By comparing the experimental results with various trackers on the above two RGB-T datasets, it can conclude that the tracker in this paper achieves better performance both in terms of precision and success rate. This proves our proposed tracker can cope with various complex environments and achieve robust tracking.

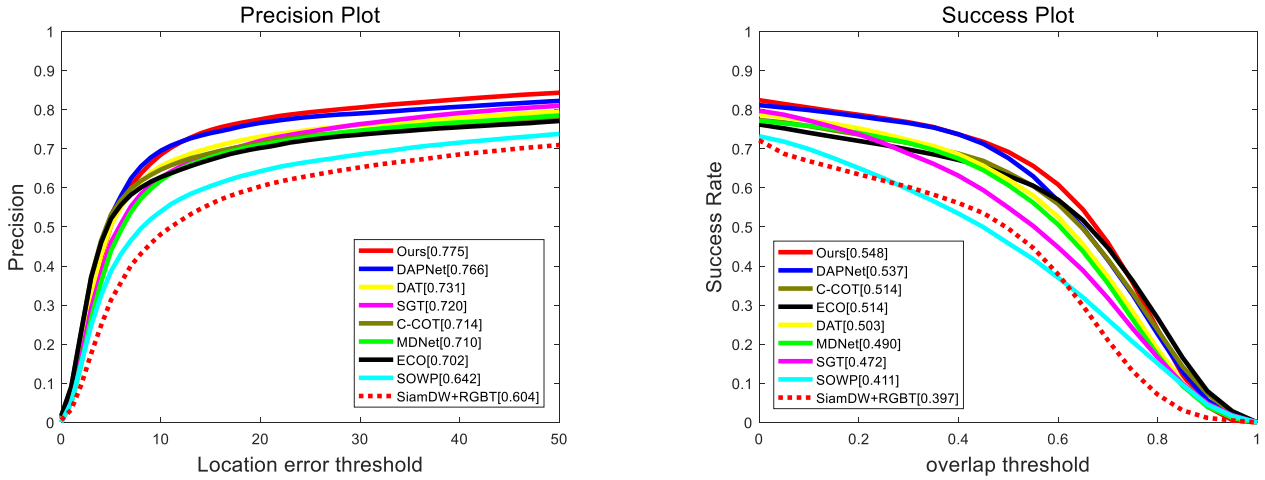

Figure 6. Precision rate (PR) and success rate (SR) curves of different tracking results on RGBT234, where the representative PR and SR scores are presented in the legend.

To explicitly show the tracking performance of our tracker, we provide three sequences for comparison, which cover different challenging attributes in the RGBT234 dataset in Figure 7. As shown in A, our method performs well. In contrast, other trackers lose the object when occlusion or camera motion occurs caused by the object moving. This indicates that our PMM and the improved positive sample selection strategy are effective, enabling our tracker to adapt to environmental changes continuously and reducing the negative impact of background information. Thereby, the robustness of object tracking can be enhanced. In B, the illumination, which is low in the object region, and the object, which has a similar temperature to other objects or the background environment, make it almost invisible in the thermal image. However, our algorithm outperforms other trackers, indicating that MWPM enables our tracker to fully use multi-modal information. In C, our algorithm can still achieve precise localization and predict the best bounding box even when the object size changes during its motion.

(**A**) (**B**) (**C**)

Figure 7. The different colored boxes in the figure correspond to different trackers: the red box is the result box of our tracker, the green box is SGT, the blue box is C-COT, and the yellow box is ECO. (**A**): Results of sequence greyman; (**B**): results of sequence woman6; (**C**): results of sequence car66.

5. Discussion

The results on the above datasets show that our tracker performs well under challenging conditions such as large-scale changes, fast motion, deformation, and small objects. These show that the combination of the pixel-matching module and the improved fully convolutional anchor-free position prediction network effectively distinguish the object from other similar objects in the scene and make up for the lack of the ability of the Siamese network to distinguish the similarity between the tracking object and the background.

In addition, the outstanding performance in low illumination, thermal crossover, and other scenarios shows that the utility of the multi-modal weight penalty module enables us to make full use of the correlation and complementarity of information between RGB and TIR modalities, which improves the quality of fusion features and improves the performance of the tracker. This is particularly important in scenarios where tracking using only RGB modality information may not be sufficient to achieve accurate tracking.

Although our tracker has achieved excellent performance in most scenarios, it has room for improvement. For example, when the target object is partially or completely occluded by other objects in the scene, the performance of the tracker is significantly reduced. This is because our tracker relies on a combination of appearance and motion cues to track the object destroyed by occlusion. We will take measures such as updating online to minimize the impact of occlusion on our tracker performance. Furthermore, our proposed method currently requires RGB-T data, and how to extend it to other modalities needs further study.

Overall, our results show that the proposed tracker is effective in dealing with various challenges encountered in object tracking. Meanwhile, the real-time running speed of 34 FPS fully meets the real-time requirements.

6. Conclusions

This paper proposes a novel high-speed, robust RGB-T tracker. A multi-modal weight penalty module is designed, which enables the new tracker to take full advantage of two modal features to cope with various lighting challenges. Combined with the proposed pixel-matching module and an improved anchor-free bounding box prediction network, the new tracker can effectively suppress the effects of cluttered backgrounds and obtain the position of objects more accurately and quickly. The new tracker achieves the advanced performance on two publicly available RGB-T benchmark datasets through the extensive experimental demonstrations.

In future work, we will continue to explore how to better integrate multi-modal information for object tracking and design more excellent algorithms dealing with the object occlusion problems.

Author Contributions: Conceptualization, D.L. and M.C.; methodology, D.L. and Y.Z.; software, Y.Z. and H.C.; validation, Y.Z.; formal analysis, D.L. and H.C.; investigation, M.C.; resources, D.L. and Y.Z.; data curation, M.C.; writing—original draft preparation, D.L. and Y.Z.; writing—review and editing, D.L. and M.C.; visualization, M.C. and H.C.; supervision, D.L. and M.C.; project administration, D.L. and M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) grant number 61801340.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, F.; Wang, X.; Zhao, Y.; Lv, S.; Niu, X. Visual object tracking: A survey. *Comput. Vis. Image Underst.* **2022**, *222*, 103508. [CrossRef]
2. Zhang, X.; Ye, P.; Leung, H.; Gong, K.; Xiao, G. Object fusion tracking based on visible and infrared images: A comprehensive review. *Inf. Fusion* **2020**, *63*, 166–187. [CrossRef]
3. 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.
4. Li, C.; Liu, L.; Lu, A.; Ji, Q.; Tang, J. Challenge-aware RGBT tracking. In Proceedings of the Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XXII. Springer International Publishing: Cham, Switzerland, 2020; pp. 222–237.
5. 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, Online, 14–19 June 2020; pp. 7064–7073.
6. Zhu, Y.; Li, C.; Tang, J.; Luo, B. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE Trans. Intell. Veh.* **2020**, *6*, 121–130. [CrossRef]
7. 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]
8. 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]
9. Guo, C.; Yang, D.; Li, C.; Song, P. Dual Siamese network for RGBT tracking via fusing predicted position maps. *Vis. Comput.* **2022**, *38*, 2555–2567. [CrossRef]
10. Zhang, T.; Liu, X.; Zhang, Q.; Han, J. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on the Siamese network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1403–1417. [CrossRef]
11. Xiao, Y.; Jing, Z.; Xiao, G.; Bo, J.; Zhang, C. A compressive tracking based on time-space Kalman fusion model. *Inf. Sci.* **2016**, *59*, 012106.
12. Xiao, G.; Yun, X.; Wu, J. A new tracking approach for visible and infrared sequences based on tracking-before-fusion. *Int. J. Dyn. Control* **2016**, *4*, 40–51. [CrossRef]
13. Zhai, S.; Shao, P.; Liang, X.; Wang, X. Fast RGB-T tracking via cross-modal correlation filters. *Neurocomputing* **2019**, *334*, 172–181. [CrossRef]
14. Yun, X.; Sun, Y.; Yang, X.; Lu, N. Discriminative fusion correlation learning for visible and infrared tracking. *Math. Probl. Eng.* **2019**, *2019*, 2437521. [CrossRef]
15. Xiong, Y.J.; Zhang, H.T.; Deng, X. RGBT Dual-modal Tracking with Weighted Discriminative Correlation Filters. *J. Signal Process.* **2020**, *36*, 1590–1597.
16. 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 ACM International Conference on Virtual Reality, Hong Kong, China, 24–26 February 2018.
17. Li, C.; Lu, A.; Zheng, A.; Tu, Z.; Tang, J. Multi-adapter RGBT tracking. In Proceedings of the 2019 IEEE International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2262–2270.
18. Zhu, Y.; Li, C.; Luo, B.; Tang, J.; Wang, X. Dense feature aggregation and pruning for RGBT tracking. In Proceedings of the ACM Multimedia Conference, Nice, France, 21–25 October 2019; pp. 465–472.
19. Lu, A.; Qian, C.; Li, C.; Tang, J.; Wang, L. Duality-gated mutual condition network for RGBT tracking. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
20. Bertinetto, L.; Jack Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; pp. 850–865.
21. 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 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4282–4291.
22. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 6667–6676.
23. Zhang, X.; Ye, P.; Xiao, G.; 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 International Conference on Information Fusion, Ottawa, ON, Canada, 2–5 July 2019.
24. Zhang, X.; Ye, P.; Qiao, D.; Zhao, J.; Peng, S.; 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]
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
26. Guo, D.; Wang, J.; Cui, Y.; He, K.; Hariharan, B.; Belongie, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 6269–6277.

27. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
28. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning (PMLR), Lille, France, 6–11 July 2015; pp. 448–456.
29. 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]
30. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
32. 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]
33. 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] [PubMed]
34. 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.
35. Pu, S.; Song, Y.; Ma, C.; Zhang, H.; Yang, M.H. Deep attentive tracking via reciprocative learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. Available online: https://proceedings.neurips.cc/paper/2018/hash/c32d9bf27a3da7ec8163957080c8628e-Abstract.html (accessed on 2 December 2018).
36. 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.
37. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4591–4600.
38. 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, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Part V. Springer International Publishing: Cham, Switzerland, 2016; pp. 472–488.
39. 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 IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 3011–3019.