electronics

Article

SCA-MMA: Spatial and Channel-Aware Multi-Modal Adaptation for Robust RGB-T Object Tracking

Run Shi [1], Chaoqun Wang [2], Gang Zhao [3] and Chunyan Xu [2,*]

[1] School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; shirun@njust.edu.cn
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; wangchaoqun@njust.edu.cn
[3] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China; zhaogang@ccnu.edu.cn
* Correspondence: xuchunyan01@gmail.com; Tel.: +86-025-84315017

Abstract: The RGB and thermal (RGB-T) object tracking task is challenging, especially with various target changes caused by deformation, abrupt motion, background clutter and occlusion. It is critical to employ the complementary nature between visual RGB and thermal infrared data. In this work, we address the RGB-T object tracking task with a novel spatial- and channel-aware multi-modal adaptation (SCA-MMA) framework, which builds an adaptive feature learning process for better mining this object-aware information in a unified network. For each type of modality information, the spatial-aware adaptation mechanism is introduced to dynamically learn the location-based characteristics of specific tracking objects at multiple convolution layers. Further, the channel-aware multi-modal adaptation mechanism is proposed to adaptively learn the feature fusion/aggregation of different modalities. In order to perform object tracking, we employ a binary classification module with two fully connected layers to predict the bounding boxes of specific targets. Comprehensive evaluations on GTOT and RGBT234 datasets demonstrate the significant superiority of our proposed SCA-MMA for robust RGB-T object tracking tasks. In particular, the precision rate (PR) and success rate (SR) on GTOT and RGBT234 datasets can reach 90.5%/73.2% and 80.2%/56.9%, significantly higher than the state-of-the-art algorithms.

Keywords: spatial- and channel-aware multi-modal adaptation; RGB-T object tracking

Citation: Shi, R.; Wang, C.; Zhao, G.; Xu, C. SCA-MMA: Spatial and Channel-Aware Multi-Modal Adaptation for Robust RGB-T Object Tracking. *Electronics* **2022**, *11*, 1820. https://doi.org/10.3390/electronics11121820

Academic Editor: Stefanos Kollias

Received: 24 March 2022
Accepted: 30 May 2022
Published: 8 June 2022

1. Introduction

Object tracking, which is an important yet challenging task in the field of computer vision, has been widely applied in video surveillance, traffic monitoring and self-driving, etc. Although general object tracking with the signal-modal RGB data source has achieved significant advances during the past few years [1–5], there are still various existing difficulties due to the challenges of low illumination, smog and darkness, etc. Meanwhile, thermal infrared data are insensitive to the lighting condition and have a strong ability to penetrate haze and smog [6], but they cannot represent targets well in good lighting conditions compared with visual images. Recently, the RGB-T object tracking problem [7–10] has received more and researchers have aimed to integrate visible and thermal infrared data for robust object tracking in the severe conditions mentioned above.

Much previous work has been devoted to RGB-T object tracking to improve the target representation with visible and thermal data [7–10]. One stream aims to extract important and expressive information from multi-modal data and then designs target descriptors for boosting the tracking performance [10–12]. For example, Li et al. chose more reliable deep feature layers to construct a target descriptor [11], learned the weights of patches cropped from multi-modal data as nodes and represent the target as a graph [10,12]. These methods only use a portion of all features to reconstruct the target descriptor, and they

omit a lot of background information, which may limit the potential tracking performance. Another stream aims to learn modality weights to achieve adaptive feature fusion for robust object tracking [8,13–15]. Early weight-based methods added [14] or concatenated [15] multi-modal information together directly. Lan et al. [7] used the max-margin principle to optimize the modality weights according to classification scores, and Zhu et al. [13] adaptively learned the modality weights via convolutional neural networks. Although these above methods consider the reliability degree in different modalities to a certain extent, they cannot well consider how to adaptively perform the feature fusion aggregation of heterogeneous modalities for achieving robust target representation.

To solve the above-mentioned problem, we propose a novel spatial- and channel-aware multi-modal adaptation (named SCA-MMA) framework for boosting the performance of the RGB-T tracking task. The SCA-MMA can not only dynamically focus on the spatial location information of specific targets, but also adaptively learn the weight of each channel based on the assumption that all the feature channels have different reliabilities [16,17]. Here, the channel weights of multi-modal representation can be adaptively learned in the process of feature aggregation. We further integrate the spatial-aware mechanism in our SCA-MMA framework, which will dynamically learn the location-based characteristics of specific tracking objects at multiple convolution layers and decrease the suppression of background for robust target representation [18,19]. In order to complete the object tracking task, we employ a binary classification with two fully connected layers to predict the bounding boxes of specific target and K branches in the training stage to learn multi-domain knowledge [4]. Extensive experiments on two public datasets demonstrate that our SCA-MMA framework can achieve state-of-the-art performance when addressing the RGB-T tracking problem. We summarize the major contributions of this work as follows.

- We propose a novel spatial- and channel-aware multi-modal adaptation (SCA-MMA) framework for robust RGB-T object tracking in an end-to-end fashion. The proposed SCA-MMA can dynamically learn the location-based characteristics of specific tracking objects and simultaneously adopt channel-aware multi-modal adaptation for better consideration of the complementarity of RGB and thermal information.
- We introduce a feature aggregation mechanism to adaptively reconstruct the target descriptor for performing RGB-T object tracking. In particular, our proposed spatial-aware mechanism can adaptively learn spatial awareness to enhance the target appearance. Furthermore, we present a channel-aware multi-modal adaptation mechanism to aggregate visual RGB and thermal infrared data, which can adaptively learn the reliable degree of each channel and then better integrate the global information.
- We evaluate the proposed SCA-MMA framework on large-scale datasets (including GTOT [8] and RGBT234 [20]). The SCA-MMA achieves 90.5%/73.2% and 80.2%/56.9% in PR/SR performance, and reaches state-of-the-art performance when compared with other RGB-T trackers [9,13,21].

2. Related Work

According to its relevance to our work, we review related work in the following two aspects: feature aggregation methods for RGB-T object tracking and multi-domain object tracking.

2.1. Feature Aggregation Methods for RGB-T Object Tracking

RGB-T object tracking, which is a sub-branch of visual object tracking, aims to aggregate visible and thermal infrared images for robust object tracking in challenging conditions such as low illumination, heavy occlusion and significant appearance changes [8,20]. Existing methods focus on robust target representation via integrating multi-modal source data [10,12,13,22]. One research stream aims to reconstruct the target descriptor via extracting effective features from multi-modal data. To perform the object tracking, Li et al. [10] proposed a weighted sparse representation regularized graph learning algorithm by constructing the specific target as a graph-based descriptor. A two-stage modality-graph

regularized manifold ranking algorithm [22] was proposed to rank all patches of multi-modal data for robust target representation. A cross-modal manifold ranking algorithm [12] was then proposed to rank cropped patches from the target while considering the heterogeneous property between different modalities and noise effects. Li et al. [11] proposed FusionNet to calculate the partial derivative of loss on channels and selected these higher parts for target representation.

Another stream aims to adaptively learn modality weights, and then concatenates them together as the target descriptor [14,15]. For example, Li et al. [8,12] regularized the modality weight via reconstruction residues, Lan et al. [7] used the max-margin principle to optimize modality weights according to classification scores, and Zhu et al. [13] adaptively learned modality weights via a convolutional neural network. To employ the temporal continuity in a video sequence, the history information was integrated to obtain fusion features by computing the adaptive weights of previous frames [23]. Tang et al. [24] proposed multiple fusion strategies from different perspectives (including pixel-level, feature-level and decision-level) to boost the performance of multi-modal object tracking in video.

2.2. Multi-Domain Object Tracking

Visual object tracking is one of the fundamental branches of computer vision and has received more and more attention in the last few decades. The pivotal branch of visual object tracking regards the object tracking problem as a one-shot binary classification task [3,4,25]. For example, Nam et al. [4] proposed a multi-domain learning framework across multiple tracking sequences in the training stage and then detected the foreground in the tracking stage. Park et al. [25] exploited the metalearning algorithm in the MDNet [4] framework, which adjusted an initial model via temporal information in tracking sequences for quick optimization in the tracking stage. Jung et al. [3] introduced the RoIAligh method to extract more accurate representations for the specific target. In [26,27], multi-domain feature representation networks have been proposed to perform information fusion across frame and event domains for improving the performance of the visual object tracking task. A semi-supervised multi-domain tracking framework [28] was proposed to learn the domain-invariant and domain-specific representations through employing an adversarial regularization. Further, a filtering-based multi-sensor data fusion technique[29] was proposed to obtain improved navigational data for unmanned surface vehicle navigation. For the radar tracking problem, the decentralized fusion of Kalman and neural filters has been proposed to deal with the multi-sensor tracking of marine targets [30]. In [31], an adaptive fusion strategy was used to integrate multiple feature cues into an observation model for improved underwater target tracking.

3. Proposed Framework

In this section, we introduce the details of the proposed SCA-MMA framework, including the network architecture and spatial-aware and channel-aware multi-modal adaptation mechanisms.

3.1. Network Architecture

The overall network architecture of the SCA-MMA model is shown in Figure 1. The SCA-MMA mainly consists of three parts: feature extraction sub-network, feature aggregation sub-network and binary classification sub-network. In particular, each feature extraction sub-network is built with three convolutional layers to extract a target representation. The spatial-aware block is employed after the front two layers to obtain spatial awareness and enhance target representation. The feature aggregation sub-network first integrates these extracted features from visible as well as thermal images, and then adaptively learns channel-wise weights via the channel-aware block and aggregates the features in terms of the channel. As in [4], the binary classification sub-network, which is adopted to distinguish the specific target and background information, has K branches after two fully

connected layers to learn multi-domain knowledge in the training stage. After finishing the multi-domain learning, the multiple branches of domain-specific layers are replaced by a single branch in the tracking stage.

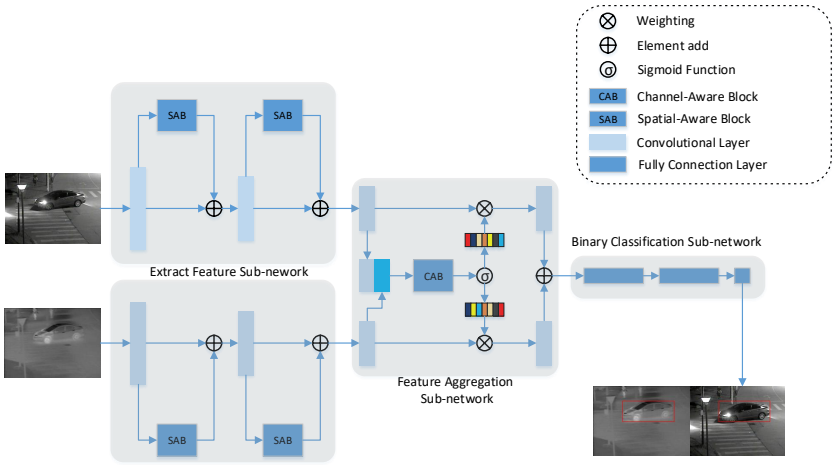

Figure 1. The overview of the proposed SCA-MMA framework for the RGB-T tracking task.

3.2. Spatial-Aware Mechanism

As shown in Figure 1, we employ the spatial-aware block in the front two convolutional layers. The details of the spatial-aware block are shown in Figure 2a. The first convolutional layer is followed by the rectified linear unit (ReLU) and local response normalization (LRN) process, and the sigmoid function is adopted after the second convolutional layer to generate spatial awareness. Here, we summarize the operations of the spatial attention block in the following equations:

$$\begin{aligned} F^1 &= LRN(ReLU(Conv(input))) \\ F^{output} &= Sigmoid(Conv(F^1)), \end{aligned} \tag{1}$$

where $Conv$, $ReLU$, LRN and $Sigmoid$ denote the convolutional layer, rectified unit, local response normalization and sigmoid function. $input$ and $output$ are the input and output of the spatial-aware block. The whole feature extraction sub-network can be summarized as follows:

$$\begin{aligned} F_1^m &= Conv(input) \\ F_2^m &= Conv(F_1^m + SAB(F_1^m)) \\ F^m &= Conv(F_2^m + SAB(F_2^m)), \end{aligned} \tag{2}$$

where SAB denotes the spatial-aware block and F^m denotes the output feature of m-th modality source data, $m \in M = \{rgb, thermal\}$ in the experiment.

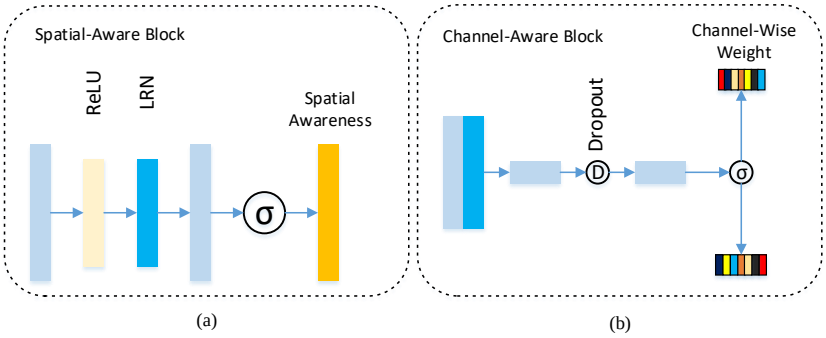

Figure 2. Diagram of spatial-aware block (**a**) and channel-aware multi-modal adaptation block (**b**).

3.3. Channel-Aware Multi-Modal Adaptation Method

The channel-aware multi-modal adaptation method can better consider heterogeneity in channel-wise weights within a single modal data. As shown in Figure 2b, the channel-aware block concatenates these extracted feature maps from two modalities. The concatenated features are fed into two fully connected layers, where each layer with 1024 output units is followed by a ReLU function. In the last layer, the dropout and softmax functions are employed in each channel dimension to obtain these channel-wise weights. Finally, under the guidance of these channel-aware weights, we can fuse these learned features for constructing a target descriptor. Here, we summarize the operations of the feature aggregation sub-network in the following equations:

$$\begin{aligned} F^{concat} &= F^R \uplus F^T \\ \{F_2^R, F_2^T\} &= fc(fc(F^{concat})) \\ \omega_R, \omega_T &= softmax(\{F_2^R, F_2^T\}, dim = channel) \\ F &= (\omega_R \otimes F_2^R) \oplus (\omega_R \otimes F_2^T) \end{aligned} \tag{3}$$

where $\uplus$, $\otimes$ and $\oplus$ denote the concatenation, channel weighting and element-wise fusion processes, $fc(\cdot)$ refers to a fully connected layer followed by ReLU as well as dropout operation, $softmax(\cdot)$ denotes the softmax function. ω_R, ω_T and F denote the learned channel-wise weights and reconstructed target descriptor by multi-modal data.

4. Experiment

4.1. Experiment Setting

We evaluate the proposed SCA-MMA framework on two large-scale benchmarks: GTOT [8] and RGBT234 [20] datasets. GTOT is an RGB-T tracking benchmark proposed by [8]. It has 50 video sequences with well-labeled visible and thermal image pairs. It is annotated with seven attributes and thus partitioned into seven subsets for analyzing the attribute-sensitive performance of RGB-T tracking approaches. RGBT234 is a large RGB-T tracking dataset, extended from the RGBT210 [10] dataset. It contains 234 video sequences, reaching approximately 23,400 frames in total and with 8000 frames for the longest video. It is annotated with 12 attributes. We use the precision rate (PR) and success rate (SR) to evaluate the quantitative performance on these two datasets. PR is the percentage of frames whose predicted location is within a threshold distance with groundtruth. SR is the percentage of frames whose overlap ratio between predicted location and groundtruth is larger than a threshold. Following the same protocols as in [8,9,13,15,20], we set the threshold to be 5 pixels for the GTOT dataset and 20 pixels for the RGBT234 dataset to evaluate PR performance. We employ the area under the curve (AUC) of the success rate as SR for quantitative performance evaluations.

The whole network is trained in an end-to-end manner. We first initialize the parameters of the convolutional layer (Conv1-Conv3) in each feature extraction sub-network using the pre-trained MDNet model [4] and randomly initialize the parameters of all the remain-

ing layers. Then, we crop positive and negative samples in training sequences randomly and minimize the cross-entropy loss by the stochastic gradient descent (SGD) algorithm, where each domain is handled separately. In the process of iteration, we randomly choose 8 frames and crop 32 positive as well as 96 negative samples in each frame to construct a minibatch in each video sequence. For positive samples, we set the IoU overlap ratio in the range 0.7∼1.0, while the negative samples are within the range 0∼0.5 IoUs. For the multi-domain learning, we set K branches for K video sequences and train the network with 100K iterations. In the front 10K iterations, we set the learning rate as 0.0001 for the feature extraction sub-network and 0.001 for the feature aggregation sub-network as well as binary classification sub-network, respectively. In the next iterations, we change the learning rate of the feature aggregation sub-network from 0.001 to 0.0001. The weight decay and momentum are fixed to 0.0005 and 0.9, respectively.

In the tracking stage, the K branches in the binary classification sub-network for multi-domain learning are replaced by a single branch for each test sequence. We then fine-tune the pretrained network in the first frame pair and update the model in subsequent frame pairs. In the fine-tuning stage, we crop 500 positive samples and 5000 negative samples with the given groundtruth bounding box. For positive samples, we set the overlap ratio in range 0.7∼1.0, while the negative samples are within the range 0∼0.5 IoUs. We fit all parameters of the feature extraction sub-network and feature aggregation sub-network. For the binary classification sub-network, we set the learning rate as 0.0001 for the front two fully connected layers and 0.001 for the last layer. We fine-tune the whole network end-to-end for 30 iterations, and train a bounding box regression model. For the given t-th frame, we crop 256 samples as candidates $\{x_t^i\}$ with the guidance of the predicted result in $t-1$-th frame, and then obtain positive scores $\{f^+(x_t^i)\}$ and negative scores $\{f^-(x_t^i)\}$. The candidate with the maximum positive score can be found as:

$$x_t^* = \arg\max_{x_t^i} f^+(x_t^i). \tag{4}$$

We find the top k candidates (i.e., k = 5). The regression technology is employed to improve target localization accuracy, and the optimal target state x^* can be seen as the mean value.

4.2. Result Comparisons

We utilize the full RGBT234 [20] dataset to construct training data and train our model for the experiment on the GTOT [8] dataset. We compare the proposed SCA-MMA with state-of-the-art trackers, including FANet [13], SGT [9], MDNet+RGBT, Struck+RGBT, L1-PF [15], ECO [32] and KCF [2]. We concatenate features used in trackers from RGB and thermal modalities as the RGB-T input of corresponding tracking algorithms [8]. Figure 3 shows that the SCA-MMA performs obviously better than the other trackers on the GTOT dataset. It gains 2.0%/3.4% in PR/SR promotion over the second-best state-of-the-art tracker. The predominant performance demonstrates that the proposed SCA-MMA can obtain the robust tracking target even in challenging conditions.

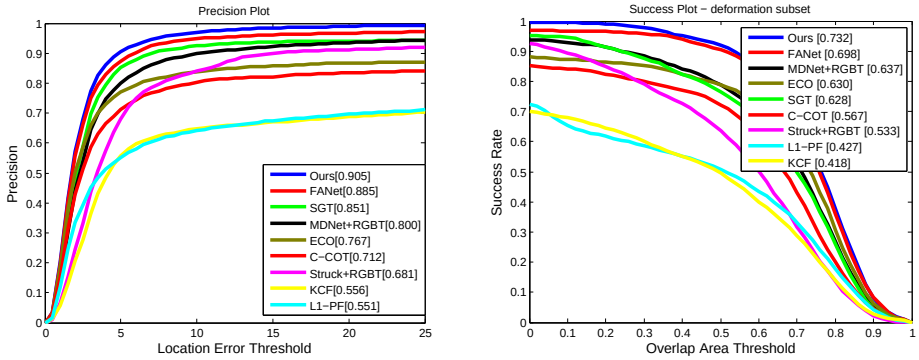

Figure 3. Performance comparisons on the GTOT dataset against state-of-the-art trackers.

We construct training data using the full GTOT dataset and train the model on the RGBT234 dataset. We compare the proposed framework with state-of-the-art trackers, including single modal trackers such as MDNet [4], ECO [32], C-COT [33], SOWP [34], SRDCF [35], CSR-DCF [36] and CFNet [37], as well as RGB-T trackers such as SGT [9], FANet [13], MDNet + RGBT, SOWP + RGBT, CSR-DCF + RGBT, L1-PF [15] and CFNet + RGBT. Here, we only display the top 12 trackers. As shown in Figure 4, the SCA-MMA framework performs the best with different evaluation metrics. Compared with the second state-of-the-art tracker, the SCA-MMA framework achieves 80.2%/56.9% in PR/SR and gains a 3.8%/3.7% improvement over the second performance tracker, as well as an 8.0%/7.4% improvement over the baseline MDNet + RGBT.

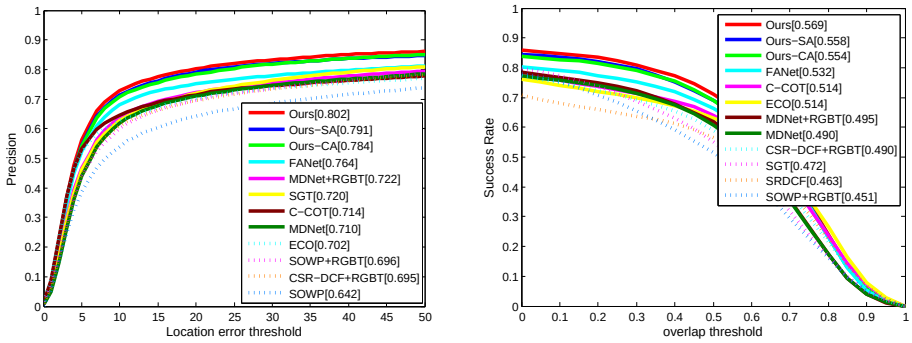

Figure 4. Performance comparisons on the RGBT234 dataset against state-of-the-art trackers.

The attribute-based results on the RGBT234 dataset are shown in Table 1. The best, second and third results are in red, green and blue colors, respectively. It contains all 12 attributes annotated on the RGBT234 dataset: no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM) and background clutter (BC). As shown in Table 1, our framework achieves the best performance in all attributes. Compared with the baseline MDNet+RGBT, we obtain over 10%/8% PR/SR improvement in DEF (deformation) and BC (background clutter) challenges with the sharp target appearance changes. It demonstrates that the proposed spatial-aware mechanism can learn spatial awareness adaptively and enhance the target information for robust object tracking. In the challenging conditions, including HO (heavy occlusion) and TC (thermal crossover), the performance of the SCA-MMA framework is significantly higher than the baseline, which demonstrates the efficiency of the proposed channel-aware multi-modal adaptation mechanism.

Table 1. Attribute-based PR/SR score(%) on the RGBT234 dataset against other RGB-T trackers.

	L1-PF [15]	CFNet + RGBT	CSR-DCF + RGBT	SOWP + RGBT	SGT [9]	MDNet + RGBT	FANet [13]	Ours
NO	56.5/37.9	76.4/56.3	82.6/60.0	86.8/53.7	87.7/55.5	86.2/61.1	84.7/61.1	90.5/66.8
PO	47.5/31.4	59.7/41.7	73.7/52.2	74.7/48.4	77.9/51.3	76.1/51.8	78.3/54.7	83.0/58.8
HO	33.2/22.2	41.7/29.0	59.3/40.9	57.0/37.9	59.2/39.4	61.9/42.1	70.8/48.1	72.9/50.7
LI	40.1/26.0	52.3/36.9	69.1/47.4	72.3/46.8	70.5/46.2	67.0/45.5	72.7/48.8	82.7/56.0
LR	46.9/27.4	55.1/36.5	72.0/47.6	72.5/46.2	75.1/47.6	75.9/51.5	74.5/50.8	80.7/55.4
TC	37.5/23.8	45.7/32.7	66.8/46.2	70.1/44.2	76.0/47.0	75.6/51.7	79.6/56.2	82.6/59.7
DEF	36.4/24.4	52.3/36.7	63.0/46.2	65.0/46.0	68.5/47.4	66.8/47.3	70.4/50.3	85.3/55.3
FM	32.0/19.6	37.6/25.0	52.9/35.8	63.7/38.7	67.7/40.2	58.6/36.3	63.3/41.7	74.3/49.3
SV	45.5/30.6	59.8/43.3	70.7/49.9	66.4/40.4	69.2/43.4	73.5/50.5	77.0/53.5	78.7/56.5
MB	28.6/20.6	35.7/27.1	58.0/42.5	63.9/42.1	64.7/43.6	65.4/46.3	67.4/48.0	69.9/51.1
CM	31.6/22.5	41.7/31.8	61.1/44.5	65.2/43.0	66.7/45.2	64.0/45.4	66.8/47.4	74.5/54.0
BC	34.2/22.0	46.3/30.8	61.8/41.0	64.7/41.9	65.8/41.8	64.4/43.2	71.0/47.8	78.3/52.7
ALL	43.1/28.7	55.1/39.0	69.5/49.0	69.6/45.1	72.0/47.2	72.2/49.5	76.4/53.2	80.2/56.9

4.3. Algorithm Analysis

Figure 5 presents the qualitative comparison of our proposed framework versus state-of-the-art RGB-T trackers on four video sequences, including SGT and MDNet+RGBT. Overall, our SCA-MMA framework is effective in handling these challenging conditions, such as low illumination, occlusion, thermal cross, deformation, background clutter and appearance change. For the elecbike10 sequence, our framework performs well in low illumination and heavy occlusion conditions, while other trackers lose the target when occlusion happens. For the fog sequence, when occlusion and bad weather happen, our framework can achieve the robust tracking target by adaptively aggregating the visible and thermal data.

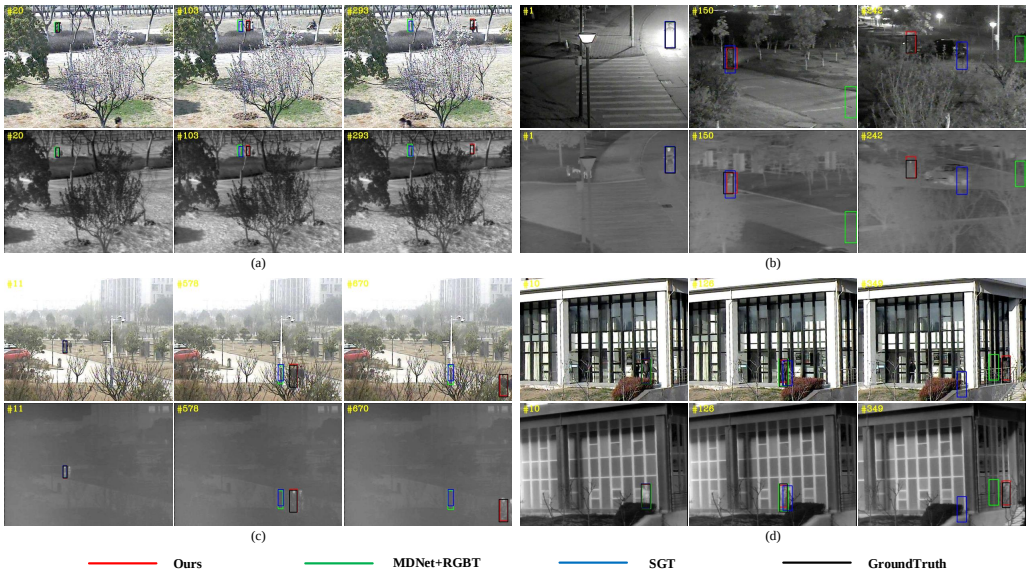

Figure 5. Qualitative performance against state-of-the-art RGB-T trackers on four video sequences. (**a**) diamond; (**b**) elecbike10; (**c**) fog; (**d**) maninglass.

To demonstrate the effectiveness of our proposed channel-aware adaptation and spatial-aware adaptation methods, we perform pruning experiments under two experimental settings on the RGBT234 dataset, including the object tracker with only with channel-aware adaptation (named "Ours-CA") and the tracker with only spatial-aware adaptation (named "Ours-SA"). The detailed performance comparisons are shown in Figure 4. The object tracker with only channel-aware adaptation can achieve 79.1% and

55.5% in terms of PR and SR, which are lower by 0.9% and 1.1% than the SCA-MMA. The tracker with only spatial-aware adaptation obtains 78.4 and 55.4% in terms of PR and SR, and its performance also reduces when compared with both "Ours-CA" and SCA-MMA methods. It clearly demonstrates that both the channel-aware adaptation and spatial-aware adaptation mechanisms can improve the performance of the RGB-T object tracking task to some extent. From the attribute-based performance shown in Table 2, we can see that in the challenges of low illumination and thermal crossover, the CA framework performs better than the SA framework. In background clutter and deformation conditions with large target appearance changes, the SA framework is far more robust than the CA framework. It demonstrates that the spatial-aware mechanism can promote target appearance in the feature extraction stage, and the channel-aware multi-modal adaptation method can handle the target reconstruction task via learning channel-wise weights in challenging conditions. Our proposed framework integrates both spatial- and channel-aware feature adaptation and achieves state-of-the-art performance.

We further employ the proposed SCA-MMA framework on the platform of Pytorch with E5-2620 V4 @2.10GHz and NVIDIA TITAN Xp. As shown in Table 3, the mean speed of our framework on the GTOT and RGBT234 datasets can reach 1.3 FPS, while the MDNet and MDNet+RGBT are 3.2 FPS and 1.6 FPS, respectively. Compared with the MDNet+RGBT tracker, the SCA-MMA framework gains an 8.0%/7.4% improvement on the RGBT234 dataset and 10.5%/9.5% improvement on the GTOT dataset with a comparable tracking speed (1.3 FPS versus 1.6 FPS).

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

	NO	**PO**	**HO**	**LI**	**LR**	**TC**	**DEF**	**FM**	**SV**	**MB**	**CM**	**BC**	**ALL**
Ours-CA	87.3/64.8	83.8/59.4	68.9/47.3	82.3/55.6	78.4/53.6	79.0/55.6	72.0/52.7	68.0/45.4	77.1/55.3	68.5/50.1	72.2/52.1	74.5/50.6	78.4/55.4
Ours-SA	88.6/64.8	84.4/59.3	69.2/48.2	80.3/54.8	78.3/53.6	76.8/54.6	84.9/54.6	71.5/46.2	78.5/55.3	69.5/50.9	72.2/52.2	76.7/52.2	79.1/55.8
Ours	90.5/66.8	83.0/58.8	72.9/50.7	82.7/56.0	80.7/55.4	82.6/59.7	85.3/55.3	74.3/49.3	78.7/56.5	69.9/51.1	74.5/54.0	78.3/52.7	80.2/56.9

Table 3. PR/SR score(%) and runtime of our framework against baseline MDNet+RGBT on GTOT and RGBT234 datasets.

		MDNet	**MDNet + RGBT**	**Ours**
GTOT	PR/SR	81.2/63.3	80.0/63.7	90.5/73.2
RGBT234	PR/SR	71.0/49.0	72.2/49.5	80.2/56.9
	FPS	3.2	1.6	1.3

5. Conclusions

In this work, we have proposed a novel spatial- and channel-aware multi-modal adaptation (SCA-MMA) framework for boosting the performance of RGB-T object tracking. In particular, we have built an adaptive and effective learning process to explore the complementarity between two heterogeneous modalities. SCA-MMA has introduced a spatial-aware mechanism to enhance the feature representations of interested objects in the spatial domain. Further, we have adaptively learned these channel-wise weights with the channel-aware multi-modal adaptation mechanism for achieving the final enhanced features of tracking targets. Extensive experiments on the RGBT234 and GTOT datasets have demonstrated that the proposed SCA-MMA has achieved the state-of-the-art performance when addressing the RGB-T tracking problem. In the future, we will focus on how to design more robust feature learning methods via the metalearning methods on the multi-modal understanding tasks, such as multi-modal object recognition, detection and object tracking.

Author Contributions: C.X.: Conceptualization, supervision and project administration, analysis data, writing and editing manuscript; R.S.: mathematical formulation, data analysis and experiment,

writing manuscript. C.W.: data analysis and interpretation, editing manuscript; G.Z.: checked the numerical results and corrected the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grants Nos. 61972204), the Natural Science Foundation of Jiangsu Province (Grant No BK20191283).

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

References

1. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
2. 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]
3. 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.
4. 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, 27–30 June 2016; pp. 4293–4302.
5. 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]
6. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]
7. Lan, X.; Ye, M.; Zhang, S.; Yuen, P.C. Robust collaborative discriminative learning for RGB-infrared tracking. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
8. 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]
9. Li, C.; Wang, X.; Zhang, L.; Tang, J.; Wu, H.; Lin, L. Weighted low-rank decomposition for robust grayscale-thermal foreground detection. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 725–738. [CrossRef]
10. 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.
11. Li, C.; Wu, X.; Zhao, N.; Cao, X.; Tang, J. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing* **2018**, *281*, 78–85. [CrossRef]
12. 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.
13. Zhu, Y.; Li, C.; Lu, Y.; Lin, L.; Luo, B.; Tang, J. FANet: Quality-Aware Feature Aggregation Network for RGB-T Tracking. *arXiv* **2018**, arXiv:1811.09855.
14. Leykin, A.; Hammoud, R. Pedestrian tracking by fusion of thermal-visible surveillance videos. *Mach. Vis. Appl.* **2010**, *21*, 587–595. [CrossRef]
15. Wu, Y.; Blasch, E.; Chen, G.; Bai, L.; Ling, H. Multiple source data fusion via sparse representation for robust visual tracking. In Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 5–8 July 2011; pp. 1–8.
16. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
17. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. *arXiv* **2019**, arXiv:1903.06586.
18. Huang, P.; Yu, G.; Lu, H.; Liu, D.; Xing, L.; Yin, Y.; Kovalchuk, N.; Xing, L.; Li, D. Attention-aware Fully Convolutional Neural Network with Convolutional Long Short-Term Memory Network for Ultrasound-Based Motion Tracking. *Med. Phys.* **2019**, *46*, 2275–2285. [CrossRef] [PubMed]
19. Su, K.; Yu, D.; Xu, Z.; Geng, X.; Wang, C. Multi-Person Pose Estimation with Enhanced Channel-wise and Spatial Information. *arXiv* **2019**, arXiv:1905.03466.
20. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *arXiv* **2018**, arXiv:1805.08982.
21. Luo, C.; Sun, B.; Yang, K.; Lu, T.; Yeh, W.C. Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. *Infrared Phys. Technol.* **2019**, *99*, 265–276. [CrossRef]
22. Li, C.; Zhu, C.; Zheng, S.; Luo, B.; Tang, J. Two-stage modality-graphs regularized manifold ranking for RGB-T tracking. *Signal Process. Image Commun.* **2018**, *68*, 207–217. [CrossRef]
23. Wang, Y.; Wei, X.; Tang, X.; Shen, H.; Zhang, H. Adaptive Fusion CNN Features for RGBT Object Tracking. *IEEE Trans. Intell. Transp. Syst.* **2021**. [CrossRef]
24. Tang, Z.; Xu, T.; Li, H.; Wu, X.J.; Zhu, X.; Kittler, J. Exploring Fusion Strategies for Accurate RGBT Visual Object Tracking. *arXiv* **2022**, arXiv:2201.08673.

25. Park, E.; Berg, A.C. Meta-tracker: Fast and robust online adaptation for visual object trackers. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 569–585.
26. Zhang, J.; Zhao, K.; Dong, B.; Fu, Y.; Wang, Y.; Yang, X.; Yin, B. Multi-domain collaborative feature representation for robust visual object tracking. *Vis. Comput.* **2021**, *37*, 2671–2683. [CrossRef]
27. Zhang, J.; Yang, X.; Fu, Y.; Wei, X.; Yin, B.; Dong, B. Object Tracking by Jointly Exploiting Frame and Event Domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13043–13052.
28. Meshgi, K.; Mirzaei, M.S. Adversarial Semi-Supervised Multi-Domain Tracking. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
29. Liu, W.; Liu, Y.; Bucknall, R. Filtering based multi-sensor data fusion algorithm for a reliable unmanned surface vehicle navigation. *J. Mar. Eng. Technol.* **2022**, 1–17. [CrossRef]
30. Stateczny, A.; Kazimierski, W. Multisensor Tracking of Marine Targets: Decentralized Fusion of Kalman and Neural Filters. *Int. J. Electron. Telecommun.* **2011**, *57*, 65–70. [CrossRef]
31. Zhang, T.; Liu, S.; He, X.; Huang, H.; Hao, K. Underwater target tracking using forward-looking sonar for autonomous underwater vehicles. *Sensors* **2019**, *20*, 102. [CrossRef] [PubMed]
32. 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.
33. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 472–488.
34. 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, 7–13 December 2015; pp. 3011–3019.
35. 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.
36. Lukezic, A.; Vojir, T.; Cehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21– 26 July 2017; pp. 6309–6318.
37. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21– 26 July 2017; pp. 2805–2813.