Multi-Modal Fusion Object Tracking Based on Fully Convolutional Siamese Network

Ke Qi
qikersa@163.com.com
School of Computer Science and Cyber Engineer, Guangzhou University
Guangzhou, China

Liji Chen*
rhichardChan@gmail.com
School of Computer Science and Cyber Engineer, Guangzhou University
Guangzhou, China

Yicong Zhou
School of Computer and Information Science, University of Macau
Macau, China

Yutao Qi
School of Computer and Information Engineering, Guangzhou Huali College
Guangzhou, China

ABSTRACT

RGBT tracking incorporates thermal infrared data to achieve more accurate visual tracking. However, the efficiency of RGBT tracking may be diminished by some bottlenecks, such as thermal crossover, illumination variation and occlusion. To address the aforementioned problems, we propose a fully-convolutional Siamese-based Multi-modal Feature Fusion Network (SiamMFF) that integrates RGB and thermal features. In our work, visible and infrared images are initially processed by the Multi-Modal Feature Fusion framework (MFF) at the search and template sides, respectively. Then, the attribute-aware fusion module is introduced to conduct feature extraction and fusion for the major challenge attributes. In particular, we design a skip connections guidance module to prevent the propagation of noise and to enrich the feature information so that we can improve the tracker's discriminative ability for modality-specific challenges. The proposed SiamMFF method has been evaluated in a great number of trials on two benchmark datasets GTOT and RGBT234, and the precision rate and success rate can reach 90.5%/73.6% and 81.2%/57.3%, respectively, demonstrating the superiority of our method over existing state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Object Tracking, RGBT, Multi-modal fusion, Siamese network

ACM Reference Format:
Ke Qi, Liji Chen, Yicong Zhou, and Yutao Qi. 2023. Multi-Modal Fusion Object Tracking Based on Fully Convolutional Siamese Network. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3590003.3590084

1 INTRODUCTION

Visual object tracking based on RGB[1, 12, 16] is an important task in computer vision. However, it is difficult for RGB trackers to obtain precise tracking results in the presence of challenging problems such as low illumination, weather changes, motion blur, fast motion, and object occlusion. Although there are ways to improve on a challenging factor[3], they may still fall short when faced with multiple challenging factors, and even greatly limit the application scope of visual tracking. Therefore, RGBT tracking [2, 7, 8, 18] has become a popular research direction in recent years.

RGBT tracking can take advantage of the complementing benefits of RGB and thermal infrared data to get better object-tracking results. Therefore, it can play a significant role in the fields of automatic driving, video surveillance, intelligent transportation and other fields. At present, most of the existing work based on RGBT is to study the fusion model[7, 18] to solve all challenges. The fusion of feature information method[7] for various challenging attributes such as illumination change and thermal cross can learn a target representation under a specific attribute with a small number of parameters. Although the fusion is innovative and has already achieved remarkable success in RGBT tracking, the utilization of modality-shared and modality-specific attribute information is still limited and like most RGBT trackers, it requires a lengthy processing period to fuse multi-modal features.

In this work, we propose a fully-convolutional Siamese-based Multi-modal Feature Fusion Network (SiamMFF) for fusing RGBT object tracking to enrich the feature information while fusing the modality-shared and modality-specific information. In particular, we design a multi-modal feature fusion module to integrate RGB and thermal features. Note that there are challenges with both shared and unique issues with RGB and RGBT data, and thus we design the multi-modal feature fusion module based on CAT[7] to address the challenges that are specific to each modality and the problems that they share. And finally, we design a skip connections guidance module to enrich the information of feature images and avoid the propagation of noisy information.

This paper's contributions are summarized below.

(1) We design a multi-modal feature fusion module to integrate visible and infrared features for RGBT object tracking. The module can extract the data that are specific to each modality and the information that they share with varying difficulties to fully use the complementing properties of visible and infrared images.

(2) We design a skip connections guidance module to prevent the propagation of noise and to enrich the feature information to improve the tracker's discriminative ability for modality-specific challenges.

*Corresponding author.

(3) Extensive experiments have been conducted on the common RGBT datasets GTOT and RGBT234, and the precision and success rate can reach 90.5%/73.6% and 81.2%/57.3%, respectively, which show the effectiveness of our method in comparison to other state-of-the-art methods.

2 RELATED WORK

2.1 RGBT Tracking

With the advancement of deep learning, RGBT tracking has gradually become more widespread. It is a challenging task that is frequently influenced by factors such as thermal crossover, scale variation, and fast motion. While most of the current RGBT tracking is to manually extract attribute features from the model, EBT[20] designs an objective measurement based on edge features to generate high-quality object proposals to improve object detection accuracy and quickly locate the objects. The current three large-scale datasets, GTOT [5], RGT210 [9], and RGBT234 [6], are also released for RGBT tracking. Meanwhile, Li et al [8] use deep learning for RGBT tracking and propose a two-stream CNN and a fusion subnetwork to extract features from two modalities respectively. CAT[7] uses parallel and hierarchical challenge-aware branches to depict how an object appearance changes under specific challenges.

Although these approaches have obtained good performance on the above datasets, there is still a lot of feature information in RGBT that has not been mined, and they are unable to fully utilize feature information in multi-modality, limiting their ability to increase performance.

2.2 Attribute-aware fusion

The essence of attribute-aware fusion is to design five attributes including illumination variation (IV), thermal crossover (TC), scale variation (SV), occlusion (OCC) and fast motion (FM) according to the five main challenges of RGBT. FM, SV, and OCC are modality-shared, and IV and TC are modality-specific. Li et al [18] propose the challenge-aware RGBT tracker for a modality-shared challenge, learning object appearance representations under different challenges. And they propose a guidance module to transmit discriminative features between modalities, which can improve some weak modalities' ability to discriminate while keeping the computational complexity low. Zhang et al [19] propose an attribute-driven representation network (ADRNet) with an attribute-driven residual branch to mine the attribute-specific characteristic and develop effective residual representations, which achieve good performance in both precisions and recall on RGBT datasets.

2.3 Fully-convolutional Siamese networks

The use of fully-convolutional networks allows online operability to stay at a fast speed. Bertinetto et al [1] propose SiamFC, using the fully-convolutional Siamese architecture for RGB tracking, in which the correlation of the two inputs is computed through bilinear layers to achieve a dense and efficient sliding window evaluation. Wang et al [16] propose SiamMask which improves the offline training procedure of fully-convolutional Siamese methods for object tracking. It adds a mask branch to the siamese network to obtain the most accurate box by directly predicting the mask of the object. Then it uses a vector to encode a response of a candidate window (RoW) of masks and perform depth-wise convolution followed by cascading 1x1 convolutions to increase the dimension and achieve efficient operation. It also proposes a top-down refine module to enhance the precision of segmentation.

3 MULTI-MODAL FUSION OBJECT TRACKING

3.1 SiamMFF Network

As shown in Fig.1, we present a fully-convolutional Siamese-based Multi-Modal Feature Fusion Network (SiamMFF) for RGBT object tracking, and the primary component of SiamMFF is the module of multi-modal feature fusion(MFF), which comprises of five attribute-specific fusion, an adaptive aggregation layer and a skip connections guidance module. In specific, we adopt SiamMask network[16] as the backbone network to allow fast speed and operability. On each side of the input, we embed MFF module to extract modality-shared and modality-specific information in visible and infrared images which is crucial for tracking. First, we input visible and thermal images and crop them to generate search images and template images centered on the tracked target object. Then the MFF module extracts modality-specific features, and the skip connections guidance module guides the fusion of modality-specific information at the same time. Next, the attributes fused by five challenge branches and the convolutional features of two template images are concatenated. The classification results and regression results are then generated by deep cross-correlation of the template and search features, and a more accurate object mask is generated following the strategy of [14], using multiple upsampling layers and a skip-connected refinement module to merge low and high-resolution features. After generating the binary mask, the Box module automatically generates bounding boxes from it using the minimum bounding rectangle strategy.

3.2 Multi-modal Feature Fusion

In GTOT[5] and RGBT234[6] datasets, the five major challenge attributes contained in each video frame are manually annotated. We design an attribute-aware fusion module based on CAT[7] to better extract and fuse the modality-shared and modality-specific branches for the tracking task. Specifically, the MFF component removes the RoIAlign layer and the fully connected layer of CAT, and only uses the convolutional layers modified from VGG-M[2] for feature extration. Among them, the three convolutional layers have a kernel size of 7×7, 5×5 and 3×3, respectively. We remove maximum pooling layers to retain more feature information in the next layer. In the second block with the padding as 2 to retain boundary information and maintain the image output size into the next layer of convolution, we modify the stride as 2 in the third block to 2 to boost the model's computational efficiency.

As shown in Fig. 2, two 255×255 pixel RGB and RBGT images are input to the MFF module. After the convolution operation of each stage, 96×125×125, 256×63×63 and 512×31×31 are output in turn, concatenating the two modalities' feature maps and output a feature map of 1024×31×31.

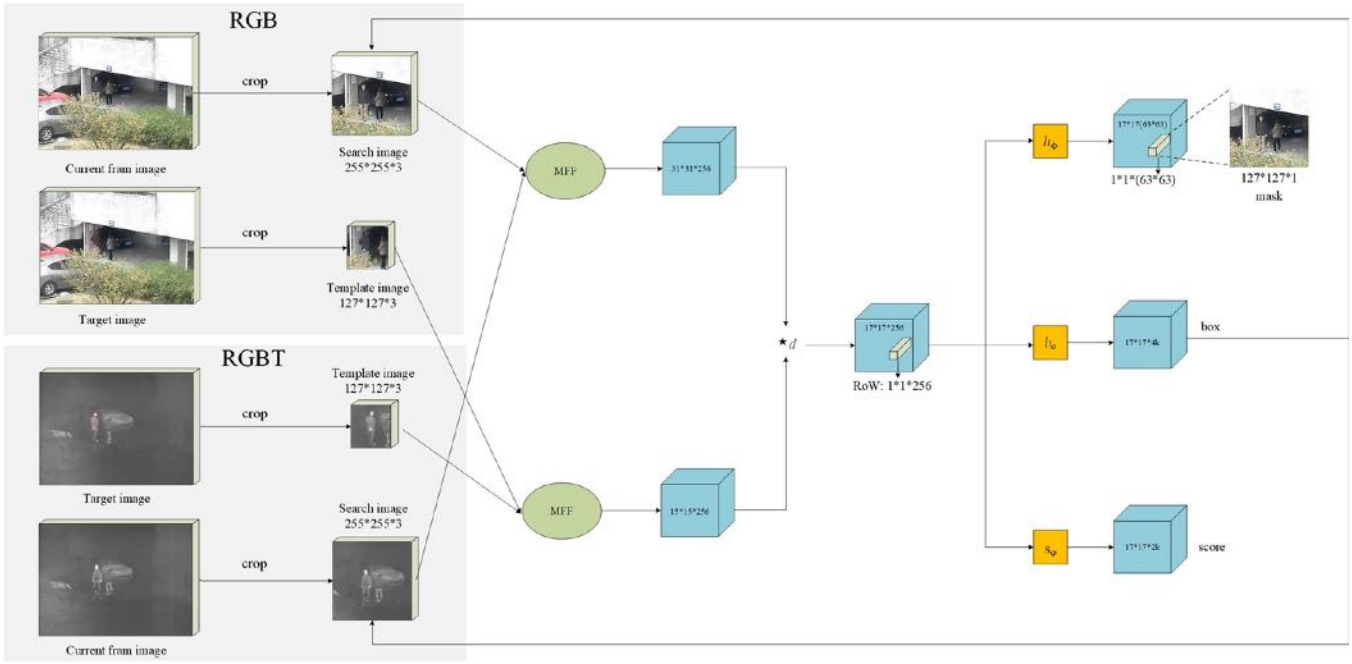

Figure 1: The structure of SiamMFF network framework. MFF represents a multi-modal feature fusion module. Mask module combines low and high resolution features through the use of several refinement modules composed of upsampling layers and skip connections. Box generation uses the minimum bounding rectangle strategy. Herein, ⋆d denotes depth-wise cross-correlation.

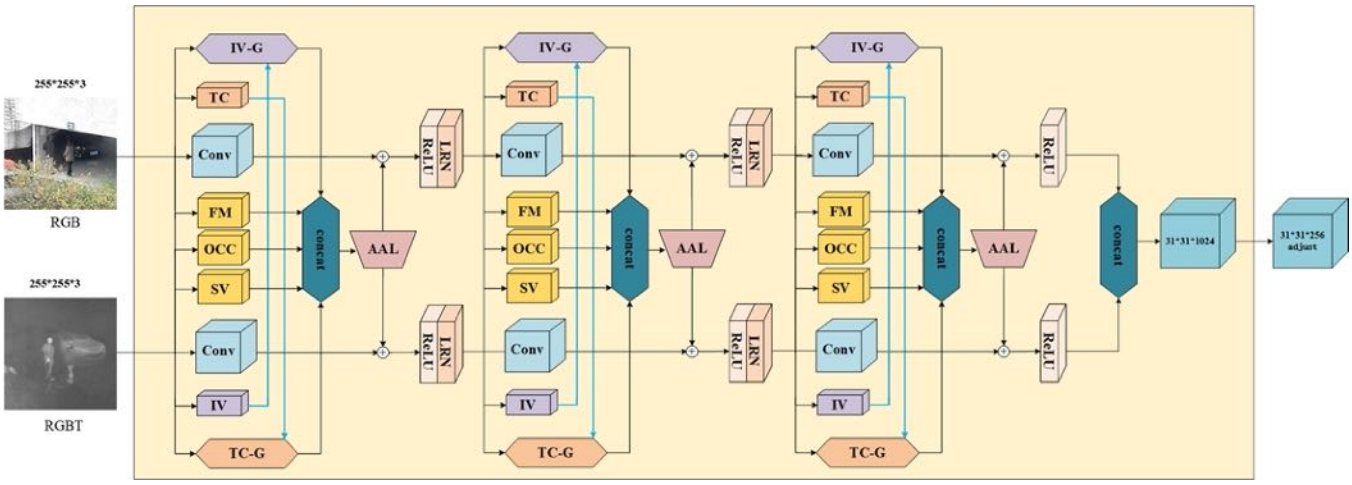

Figure 2: The structure of Multi-Modal Feature Fusion model (MFF). Herein, + represents an element-wise addition operation. AAL stands for adaptive aggregation layer. FM, OCC and SV are abbreviations for fast motion, occlusion, and scale variation. IV-G and TC-G are illumination variation and thermal crossover with the guidance module.

3.3 Skip connections guidance module

For TC and IV challenges, we use the skip connections guidance module to improve the discrimination of guided modality during tracing. As shown in Fig.3, we design the skip connections guidance module motivated by the guidance module on CAT[7]. Unlike only using feature shift in CAT, we use skip connections, which are commonly used to enrich image details, such as U-Net[15], and the results of the studies support our design's efficacy. Additionally, in order to avoid noisy information, we introduce a gate mechanism. Specifically, we use a convolutional layer with 1 x 1 kernel size and in order to learn a nonlinear mapping, we use a layer of nonlinear

activation. The gate operation is accomplished using element-wise sigmoid activation.

Here is the formulation of our skip connections module:

$$\alpha = \omega_1 * \mathbf{x} + b_1$$
$$\beta = \omega_2 * \mathbf{z} + b_2$$
$$\beta = \alpha + \beta$$
$$\theta = \omega_3 * ReLU(\beta) + b_3$$
$$\tilde{\theta} = \sigma(\theta) * \beta$$
$$\mathbf{z} = \mathbf{z} * \tilde{\theta}$$

The convolutional layer's weight and bias are denoted by ω_i and b_i. σ denotes the sigmoid function. Point-by-point feature conversion without gate operation is represented by α and β, and $\tilde{\theta}$ represents point-by-point feature conversion with gate operation. The feature maps of the preceding and guided modalities are indicated, respectively, by $\mathbf{x}$ and $\mathbf{z}$.

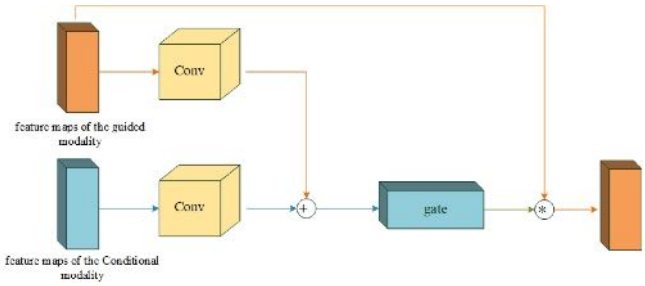

Figure 3: Skip connections Guidance modules.

4 EXPERIMENTS

4.1 Evaluation Data

We put our SiamMFF to the test on two RGBT datasets., i.e., GTOT[5] and RGBT234[6]. The GTOT dataset comes from the research group of Chenglong Li of Anhui University which consists of 50 statistically biased video sequences. The sequences of grayscale image-thermal infrared image pairs are divided into 7 subsets according to different properties. The RGBT234 dataset includes 234 RGBT video sequence pairs and their corresponding ground truth values. There are 234K frames in all. There are 12 attributes in the video sequence annotation, which are useful for evaluating the efficacy of various tracking algorithms for various challenging attributes.

4.2 Evaluation metrics

On the GTOT and RGBT234 datasets, our evaluation metrics use success rate (SR) and precision rate (PR) via one-pass evaluation. The fraction of successfully tracked frames with overlaps greater than thresholds is measured by SR. PR is the proportion of all frames in which the true distance between the tracking result's center point and the ground truth is smaller than the threshold. Since GTOT dataset contain mostly small objects, the threshold is set to 5 pixels on GTOT and 20 pixels on RGBT234.

4.3 Quantitative Comparison

We put our SiamMFF to the test on two benchmark datasets, GTOT and RGBT234. We train the challenge-aware branches in our multi-modal feature fusion module with corresponding challenge-based training data collected from the RGBT234 dataset by attribute labels in the GTOT dataset test. Then, we train MFF using the whole RGBT234 dataset. The training dataset for RGBT234 testing is GTOT, and the whole training procedure is similar to what we have described above.

To evaluate the usefulness of our SiamMFF, we run it through two RGBT datasets and compare its performance to that of numerous state-of-the-art trackers, such as ADRNet[19], MANet++[11], CAT[7], MANet[10], DAFNet[4], mfDiMP[18], DAPNet[21], SGT[9], FANet[22], MaCNet[17], RT-MDNet+RGBT [12] and MDNet[13]+RGBT.

Fig.4 depicts the comparison findings on GTOT and RGBT234 datasets. As shown in Fig.4(a), our SiamMFF surpasses CAT by 1.6%/1.9% in PR/SR which also use attribute-aware fusion. We also achieve comparable results when compared with the state-of-the-art approach ADRNet, where SiamMFF is 0.3% lower in SR but 0.1% higher in PR. And as shown in Fig.4(b), SiamMFF achieves the best tracking performance on RGBT234 dataset. Compared with ADRNet, which is the top advanced tracker in RGBT234, our PR/SR is 0.5%/0.3% higher than it and when compared with CAT, we advance them by 0.8%/1.2% in PR/SR. Furthermore, we outperform MDNet+RGBT in PR/SR by more than 9%/7.8%. These findings totally support the efficacy of our strategy.

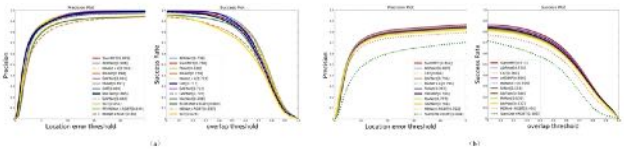

Figure 4: The assessment curve on GTOT and RGBT234 datasets. The legend displays the PR/SR representative scores.(a)GTOT, (b)RGBT234

4.4 Ablation Study

We conduct an ablation study on GTOT and RGBT234 to validate the efficiency of the major components of our approach. Aim to test the effectiveness of the proposed network, we implement the two special variants of our method described below: 1. SiamMFF-AAF, all attribute-aware branches are combined by elements addition, deleting the structure of the Attribute-aware fusion module, in order to test the efficacy of the proposed modal feature fusion approach. 2. APFNet-SCGM, the special challenge branches are fused by elements addition, removing the skip connections guidance modules to verify the effectiveness of the proposed skip connections guidance module. Table1 shows the experimental findings on the GTOT and RGBT234 datasets. On the GTOT and RGBT234 datasets, we take SiamMFF without the skip connections guidance module as the baseline and compare CAT's guidance module with our skip connections module on our SiamMFF model to validate the proposed method's efficacy. As shown in Table 2, we can see our

skip connections guidance module outperforms CAT on the two datasets.

Table 1: The PR/SR scores of various versions induced by our SiamMFF on two RGBT benchmark datasets are used to verify the usefulness of our model.

		SiamMFF-AAF	SiamMFF-SCGM	SiamMFF
GTOT	PR	0.853	0.861	**0.905**
	SR	0.685	0.720	**0.736**
RGBT234	PR	0.755	0.793	**0.812**
	SR	0.501	0.561	**0.573**

Table 2: Compare the performance of the skip connections guidance module and CAT's guidance module on SiamMFF on two RGBT benchmark datasets.

		Baseline	SiamMFF-CAT	SiamMFF
GTOT	PR	0.861	0.884	**0.905**
	SR	0.720	0.725	**0.736**
RGBT234	PR	0.793	0.798	**0.812**
	SR	0.561	0.567	**0.573**

4.5 Qualitative Results

Fig. 5 displays the qualitative results of 7 different trackers on two sequences under different challenging circumstances, such as scale variation(e.g. Otcbvs) and thermal crossover(e.g. Motorbike). We can see that our SiamMFF is more robust under these challenging situations and gets better results under the qualitative comparison of bounding boxes, which demonstrates the effectiveness of our method.

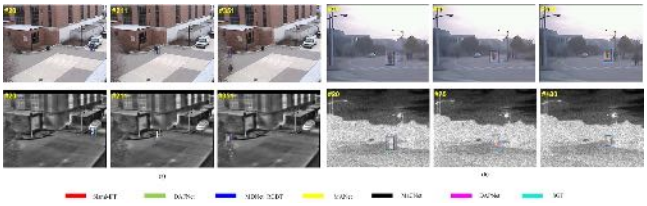

Figure 5: Qualitative results of 7 trackers on two sequences in the GTOT dataset. (a) Otcbvs,(b) Motorbike.

5 CONCLUSION

In this paper, we propose a fully-convolutional Siamese-based Multi-Modal Feature Fusion Network (SiamMFF) to fully integrate multi-modal feature information and introduce the attribute-aware module into the modified fully-convolution Siamese network SiamMask, to enhance the fusion at the attribute feature level. We also design a skip connections guidance module to enrich feature image information. The proposed algorithm performs well when evaluated on commonly used benchmark datasets and compared with mainstream algorithms. The experimental results validate the proposed method's effectiveness and feasibility. In further exploration, we plan to try more fusion structures under challenging circumstances such as motion blur and size change so that the model can fully fuse the information between multi-modalities. We will also further consider adding visual reasoning of optical flow direction under the presumption of guaranteeing the model runs light.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*. Springer, 850–865.

[2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

[3] Siyu Di and Wensheng Sun. 2022. Research on Low Illumination Image Processing Algorithm Based on Adaptive Parameter Homomorphic Filtering. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. 681–685. https://doi.org/10.1109/CACML55074.2022.00118

[4] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. 2019. Deep adaptive fusion network for high performance RGBT tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[5] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing* 25, 12 (2016), 5743–5756.

[6] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition* 96 (2019), 106977.

[7] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. 2020. Challenge-aware RGBT tracking. In *European Conference on Computer Vision*. Springer, 222–237.

[8] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing* 281 (2018), 78–85.

[9] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*. 1856–1864.

[10] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. 2019. Multi-adapter RGBT tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[11] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. 2021. RGBT tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing* 30 (2021), 5613–5625.

[12] Yuwei Lu, Yuan Yuan, and Qi Wang. 2020. A dense connection based network for real-time object tracking. *Neurocomputing* 410 (2020), 229–236.

[13] Hyeonseob Nam and Bohyung Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4293–4302.

[14] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. 2016. Learning to refine object segments. In *European conference on computer vision*. Springer, 75–91.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[16] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 1328–1338.

[17] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. 2020. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors* 20, 2 (2020), 393.

[18] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. 2019. Multi-modal fusion for end-to-end rgb-t tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[19] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 2021. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *International Journal of Computer Vision* 129, 9 (2021), 2714–2729.

[20] Gao Zhu, Fatih Porikli, and Hongdong Li. 2016. Beyond local search: Tracking objects everywhere with instance-specific proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 943–951.

[21] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. 2019. Dense feature aggregation and pruning for rgbt tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 465–472.

[22] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. 2020. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE Transactions on Intelligent Vehicles* 6, 1 (2020), 121–130.