SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Learning cross-modal interaction for RGB-T tracking

Chunyan Xu, Zhen Cui*, Chaoqun Wang, Chuanwei Zhou & Jian Yang

Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094*, China*

Appendix A Related work

Visual object tracking: Visual object tracking, which is an important but challenging problem in the field of computer vision, has attracted increasing attention owing to its broad application in vision understanding. Correlation filter-based tracking methods [2, 11,13,19] have shown advantages because of their high computational efficiency with the use of fast Fourier transforms. For example, a classic type of correlation filter, called the minimum output sum of squared error (MOSSE) filter [2], was first proposed to produce stable correlation filters when initialized from the first frame. Some extensions of the correlation filter trackers were subsequently proposed. Using the well-established theory of circulant matrices, the CSK tracker [20] established a link to Fourier analysis that opened up the possibility of extremely fast learning and detection using the fast Fourier transform (FFT). To incorporate color information in the process of visual tracking, the CSK tracker was extended to multidimensional color features by defining an appropriate kernel [12]. To tackle the problem of the fixed template size in kernel correlation filter tracker, Li et al. [33] extended the correlation filter-based tracker and then proposed an effective scale adaptive scheme to improve the tracking capability. To achieve high-speed tracking, a kernelized correlation filter (KCF) [19] was presented to leverage the powerful kernel trick at the same computational complexity as that of linear correlation filters. In contrast to the conventional discriminative correlation filters framework, Danelljan et al. [13] employed an implicit interpolation model to learn convolution filters in a continuous spatial domain. In [11], a discriminative scale space tracker (DSST) was proposed to learn separate correlation filters for explicit translation and scale estimation.

Convolutional neural networks (CNNs) with powerful feature representation abilities have facilitated significant advances in computer vision tasks (e.g., whole-image classification [18, 37, 42], semantic segmentation [4, 5, 17], and object detection [15]). In addition, numerous handcrafted features have been replaced by deep CNN features to represent the target appearance in the task of visual object tracking [9, 10]. Wang et al. [40] trained a stacked denoising auto-encoder offline to learn generic image features by employing auxiliary natural images and then transferred the learned knowledge from offline training to the online tracking process. An extended version of DLT, named SO-DLT [39], was proposed to pre-train the CNN to recognize objects and generate a probability map to better fit the characteristics of object tracking. Nam et al. [35] proposed a multi-domain convolutional neural network (MDNet) learning framework to address the problem of visual tracking, where a CNN pre-trained by multi-domain learning is updated online in the context of a new sequence to learn domain-specific information adaptively. Inspired by MDNet and fast R-CNN [16], a real-time visual tracking algorithm [21] was proposed to accelerate the feature extraction procedure and learn more discriminative models for instance classification. Li et al. [32] presented a correlation filter neural network (CFNN) tracker by employing the advantages of both deep learning and correlation filter-based methods. Efficient convolution operators (ECO) [9] were proposed to dramatically reduce the number of parameters in the DCF model and simultaneously improve the tracking speed and robustness. In [47], deeper and wider Siamese network architectures were designed for achieving real-time visual tracking based on the no-padding residual units. Deformable Siamese attention networks [44] were proposed to handle object tracking by employing a Siamese attention mechanism, that is, deformable self-attention and cross-attention. Zhang et al. [48] proposed a tracking framework by combining the anchor-free network with an efficient feature combination module, which used a 2D spatial transformation to align the feature sampling locations with predicted bounding boxes. The transformer tracking work [6] presented an attention-based feature fusion network that effectively combines the template and search region features solely using attention without correlation. The spatio-temporal transformer tracking method [43] was then proposed to capture global feature dependencies of both spatial and temporal information in video sequences. Despite breakthroughs in visual object tracking, many challenges still exist; in particular, tracking target objects under conditions of low illumination leads to loss of essential object information.

RGB-T object tracking: Owing to the special advantages and accessibility of thermal infrared cameras, RGB-T object tracking has recently has gained increasing attention [27, 29, 45, 50]. According to different fusion methods of multi-modal data information, existing RGB-T tracking methods are mainly based on weight learning [31,50], sparse representation [25,29], and deep learning methods [38, 44].

Multi-modal weight learning algorithms obtain the weights/confidences of multi-modal data under different conditions and then fuse the multi-modal data information through the learned weights to improve multi-modal object tracking. For instance, a deep quality-aware feature aggregation network (FANet) [50] was proposed to achieve quality-aware aggregation of both hierarchical features and multi-modal information for robust online RGB-T tracking. To achieve effective multimodal fusion, soft cross-modality consistency was used to enforce ranking consistency between RGB and thermal modalities while allowing sparse inconsistency [31]. By employing the sparse representation strategy, Li et al. [25] jointly optimized the sparse codes and reliable weights of different modalities to simultaneously integrate grayscale and thermal information. A grayscale-thermal object tracking method in a Bayesian filtering framework was proposed to pursue multitask Laplacian sparse representation to adaptively leverage multimodal visual data [28]. A weighted sparse representation regularized graph model [29] was also introduced to learn a robust object representation using multi-spectral (RGB and thermal) data for visual tracking. A discriminative learning framework was proposed to adaptively

* Corresponding author (email: zhen.cui@njust.edu.cn)

Table B1 Attribute-based PR score (%) on the RGBT234 dataset against state-of-the-art trackers. The best, second and third performances are represented in red, green and blue, respectively.

Method	SGT	FANet	DAPNet	DAFNet	MaCNet	C-COT	CMPP	SOWP+RGBT	MDNet+RGBT	Ours
NO	87.7	84.7	90.0	90.9	92.7	88.8	95.6	86.8	89.5	94.3
PO	77.9	78.3	82.1	85.9	81.1	74.1	85.5	74.7	79.6	87.9
HO	59.2	70.8	66.0	68.6	70.9	60.9	73.2	57.0	67.0	73.4
LI	70.5	72.7	77.5	81.2	77.7	64.8	86.2	72.3	74.5	85.2
LR	75.1	74.5	75.0	81.8	78.3	73.1	86.5	72.5	75.8	84.2
TC	76.0	79.6	76.8	81.8	77.0	84.0	83.5	70.1	73.9	79.4
DEF	68.5	70.4	71.7	74.1	73.1	63.4	75.0	65.0	70.8	77.6
FM	67.7	63.3	67.0	74.0	72.8	62.8	78.6	63.7	66.4	79.1
SV	69.2	77.0	78.0	79.1	78.7	76.2	81.5	66.4	76.8	82.8
MB	64.7	67.4	65.3	70.8	71.6	67.3	75.4	63.9	67.8	74.3
CM	66.7	66.8	66.8	72.3	71.7	65.9	75.6	65.2	71.0	76.9
BC	62.8	71.0	71.7	79.1	77.8	59.1	83.2	64.7	74.4	83.9
ALL	72.0	76.4	76.6	79.6	79.0	71.4	82.3	69.6	76.3	83.1

and collaboratively learn the classifiers and reliability weights of different modalities for RGB-infrared tracking [24]. Recently, deep learning methods, particularly convolutional neural networks, have further improved the performance and robustness of RGB-T tracking tasks. To perform the deep RGB-T tracking task, a dense feature aggregation and pruning network (DAPNet) [49] was constructed to perform an effective information fusion of different modalities in an end-to-end fashion. A multiadapter convolutional network (MANet) [27] was employed to jointly perform modality-shared, modality-specific, and instance-aware feature learning in an end-to-end trained deep framework for RGB-T tracking. A modal-aware attention network [45] was proposed to perceive the importance of each modality and to guide the adaptive fusion of dual-modality features on multiple feature layers. Wang et al. [38] proposed a cross-modal pattern-propagation (CMPP) tracking framework to diffuse instance patterns across RGB-T data in both spatial and temporal domains. In departure from the existing RGB-T tracking approaches, we attempt to mine the intrinsic correlations between two heterogeneous modalities to promote information propagation from one modality to another modality.

Appendix B Experiments

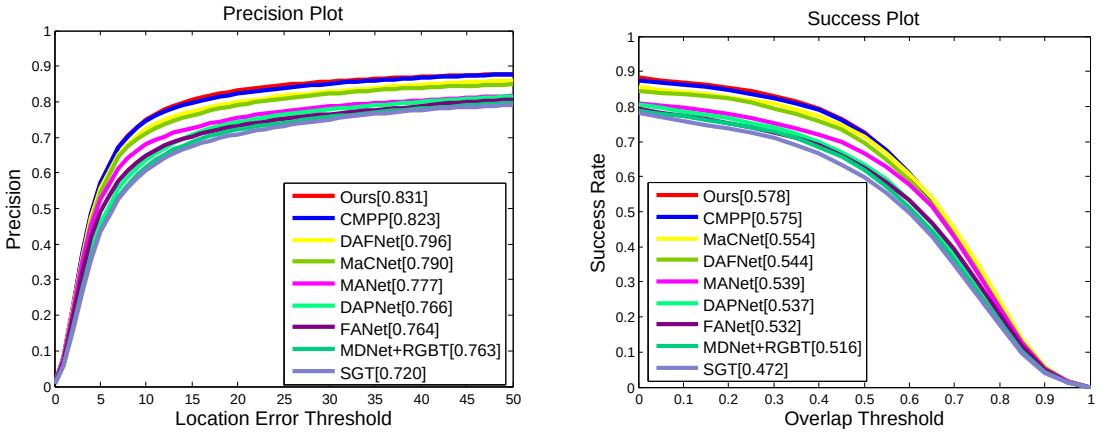
Experimental Settings: We evaluated the effectiveness of our CMI on two larger RGB-T object tracking datasets: GTOT [25] and RGBT234 [26]. The GTOT dataset [25] consisted of 50 spatially and temporally aligned pairs of RGB and thermal infrared video sequences in different environments. To comprehensively evaluate different tracking algorithms, the challenges were divided into seven categories based on the weather and time of the shoot and the status of the target. The RGBT234 dataset [26] was an extension of the RGBT210 dataset [30]. It contained 234 video pairs that were strictly aligned using two modalities. Its total frames were approximately 234K and the frame of the largest video pair was 8K. Various occlusion levels (including no, partial, and heavy occlusions) were annotated for the occlusion-sensitive evaluation of the different algorithms. For the attribute-sensitive performance analysis, the attributes of each video sequences were also annotated, including 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 movement (CM), and background clutter (BC). We employed widely used tracking evaluation metrics, including the success rate (SR) and precision rate (PR), for the quantitative performance evaluation of RGB-T object tracking. SR is the percentage of frames whose overlap ratio between the estimated bounding box and the ground truth bounding box is larger than a specified threshold. We computed the representative SR score using the area under the curves. PR is the percentage of frames whose predicted bounding box is within a threshold distance of the ground truth. Because the target objects in GTOT are small, we computed the representative PR score by setting the thresholds to 5 and 20 pixels for GTOT and RGBT234 datasets, respectively. The VOT-RGBT2019 dataset [23] contains 60 testing sequences. The targets are annotated as rotated rectangles to enable a more thorough localization accuracy. By following the standard evaluation protocol in [23], we used the accuracy, robustness, and expected average overlap (EAO) to evaluate the performance of different trackers. Accuracy and robustness reflect the accuracy and robustness of the tracker, whereas the expected average overlap reflects the overall performance of the tracker. Higher values of accuracy, robustness, and expected average overlap are desirable for a tracker.

The entire CMI network was optimized in an end-to-end manner. In the training process, the VGG-M model [3] pre-trained on the ImageNet dataset was adopted to initialize the first three convolutional layers, and other network parameters were initialized randomly. In the CMI module, we simultaneously employed two CMI learning strategies to obtain fusion features, which were then fed into the binary classification network. Positive and negative samples were randomly selected from the training video sequences. To better capture this multi-domain information, the last fully connected layer was followed by K network branches in the binary classification network, and each video sequence was used to optimize one network branch independently. In the iterative learning process, we randomly selected eight frames from each video sequence and then cropped 32 positive samples and 96 negative samples from each frame to optimize the parameters of the CMI network. A selected bounding box is considered a positive sample if it overlaps by more than 70% with a ground-truth bounding box, and a bounding box is judged to be a negative sample if it overlaps by less than 50% with a ground-truth bounding box. The parameter learning rate of the feature extraction and fusion network modules was set to 0.0001, and the learning rate of the other network parameters was 0.001. We trained our CMI network using the stochastic gradient descent method with 1000 iterations, a momentum of 0.9, and a weight decay of 0.0005. Following the same protocol in [38], we selected the entire GTOT dataset as the training set when evaluating the RGBT234 dataset and randomly selected 70 videos of the RGBT234 dataset as the training set when evaluating the GTOT dataset.

In the test stage, the K network branches used for multi-domain learning were replaced by one branch in the last layer of the binary classification network. We sampled 500 positive and 5000 negative samples from the initial frame to fine-tune the parameters of our network model. The network parameters of the feature extraction and fusion parts were fixed during the test process. In the binary classification subnetwork, the learning rate of the first two network layers was set to 0.0001, and the learning rate of the last fully connected layer was 0.001. The end-to-end CMI network could then be optimized over 30 iterations using the random gradient

Table B2 Attribute-based SR score (%) on the RGBT234 dataset against state-of-the-art trackers. The best, second and third performances are represented in red, green and blue, respectively.

Method	SGT	FANet	DAPNet	DAFNet	MaCNet	C-COT	CMPP	SOWP+RGBT	MDNet+RGBT	Ours
NO	55.5	61.1	64.4	63.6	66.5	65.6	67.8	53.7	62.6	67.4
PO	51.3	54.7	57.4	58.8	57.2	54.1	60.1	48.4	53.5	60.9
HO	39.4	48.1	45.7	45.9	48.8	42.7	50.3	37.9	44.9	50.3
LI	46.2	48.8	53.0	54.2	52.7	45.4	58.4	46.8	48.1	57.1
LR	47.6	50.8	51.0	53.8	52.3	49.4	57.1	46.2	48.9	55.3
TC	47.0	56.2	54.3	58.3	56.3	61.0	58.3	44.2	51.1	55.6
DEF	47.4	50.3	51.8	51.5	51.4	46.3	54.1	46.0	50.0	55.3
FM	40.2	41.7	44.3	46.5	47.1	41.8	50.8	38.7	43.3	49.9
SV	43.4	53.5	54.2	54.4	56.1	56.2	57.2	40.4	52.0	58.0
MB	43.6	48.0	46.7	50.0	52.5	49.5	54.1	42.1	48.0	53.2
CM	45.2	47.4	47.4	50.6	51.7	47.3	54.1	43.0	50.1	54.5
BC	41.8	47.8	48.4	49.3	50.1	39.3	53.8	41.9	48.9	54.9
ALL	47.2	53.2	53.7	54.4	55.4	51.4	57.5	45.1	51.6	57.8



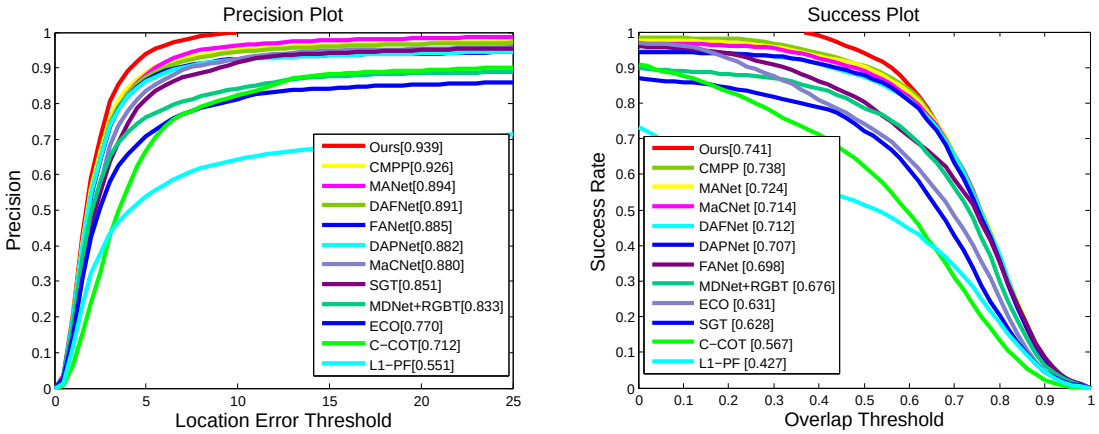
(a) PR score (b) SR score

Figure B1 Overall performance comparison with state-of-the-art methods on the RGBT234 dataset.

descent method. We employed the above learned CMI network to obtain these multi-modal features and predict the location results of the target of interest. For the tracking process in the t-frame, we sampled 256 proposal regions with a Gaussian distribution model by employing the tracking result of the $(t-1)$-th frame and obtained the positive and negative scores of all selected samples with the learned CMI network. The top-k proposal regions of the positive scores (i.e., k=5) was used to obtain the bounding boxes, and its mean value was the final predicted bounding box of the tracking target.

Results and comparisons To comprehensively evaluate the effectiveness of the proposed CMI, we compared it with several state-of-the-art approaches [14, 29, 38, 45, 50] on two public RGB-T tracking datasets. Tables B1 and B2 report the attribute-based PR/SR scores of the proposed CMI and comparisons with several state-of-the-art methods on the RGBT234 dataset, including SGT [29], FANet [50], DAPNet [49], DAFNet [14], MaCNet [45], CMPP [38], C-COT [13], SOWP [22], and MDNet [35]. Our CMI can significantly outperform existing baselines, achieving improvements in 11.1%/10.6% over SGT [29], 6.7%/4.6% over FANet [50], 6.5%/4.1% over DAPNet [49], 3.5%/3.4% over DAFNet [14], 4.1%/2.4% over MaCNet [45], 0.8%/0.3% over CMPP [38], and 6.8%/6.2% over the MDNet+RGBT method in terms of PR/SR score. Our CMI exhibits much higher performance than other existing trackers. Furthermore, when predicting object locations in different attribute-based video sequences, the PR/SR results of our CMI method achieved significant gain under most conditions, for example, 83.9%/54.9% *vs.* 62.8%/41.8% [29] for background clutter (BC), 82.8%/58.0% *vs.* 78.0%/54.2% [49] for scale variation (SV), and 77.6%/55.3% *vs.* 74.1%/51.5% [14] for deformation (DEF). In particular, our CMI obtained the best tracking results in the partial occlusion (PO) and heavy occlusion (HO) cases, which demonstrates that the proposed cross-modal fusion method can improve the RGB-T tracking performance to some extent. When the tracking objects have significant appearance changes (such as scale variation and deformation) or are moving in the complicated environments of background clutter and camera movement, we can also achieve the best tracking performance. These attribute-based results indicate that CMI performs well in various appearance changes and challenging situations. Furthermore, Fig. B1 presents a comparison of the PR/SR curves of our CMI and other state-of-the-art RGB-T tracking methods. From the above results, we can observe that our CMI performs better than the compared baselines [14, 29, 35, 38, 45, 49, 50] in both the PR and SR metrics. These above improvements may be due to the complementarity between the heterogeneous data sources, which allows the moving objects to be well located during the tracking process in the RGB-T video sequences.

Fig. B2 presents a comparison of the PR/SR curves of our CMI and other baselines on the GTOT dataset, including MANet [27], ECO [9], C-COT [13], DAFNet [14], L1-PF [41], SGT [29], FANet [50], CMPP [38], FANet [50], MaCNet [45], MDNet [35]+RGBT and SGT [29]. The proposed CMI method obtained PR and SR scores of 93.9% and 74.1%, respectively. Comparisons with previous methods [9, 27, 35, 50] demonstrate that the proposed CMI substantially outperforms them, achieving improvements of 4.5% over MANet [27], 5.4% over FANet [50], and 5.9% over MaCNet [45] in terms of the PR score, as well as improvements of 2.9% over



(a) PR score (b) SR score

Figure B2 Overall performance comparison with state-of-the-art methods on the GTOT dataset.

Table B3 Tracking results of different trackers on the VOT-RGBT2019 dataset. The best, second and third performances are represented in red, green and blue, respectively.

Method	Accuracy	Robustness	EAO
TFNet [51]	0.462	0.594	0.288
FANet [50]	0.472	0.508	0.247
MANet++ [34]	0.509	0.538	0.272
SGT [29]	0.518	0.723	0.297
MANet [27]	0.582	0.701	0.346
ATOM [8]	0.587	0.695	0.321
DiMP [1]	0.601	0.709	0.327
SiamBAN [7]	0.622	0.706	0.333
ADRNet [46]	0.622	0.766	0.396
Ours CMI	0.625	0.716	0.381

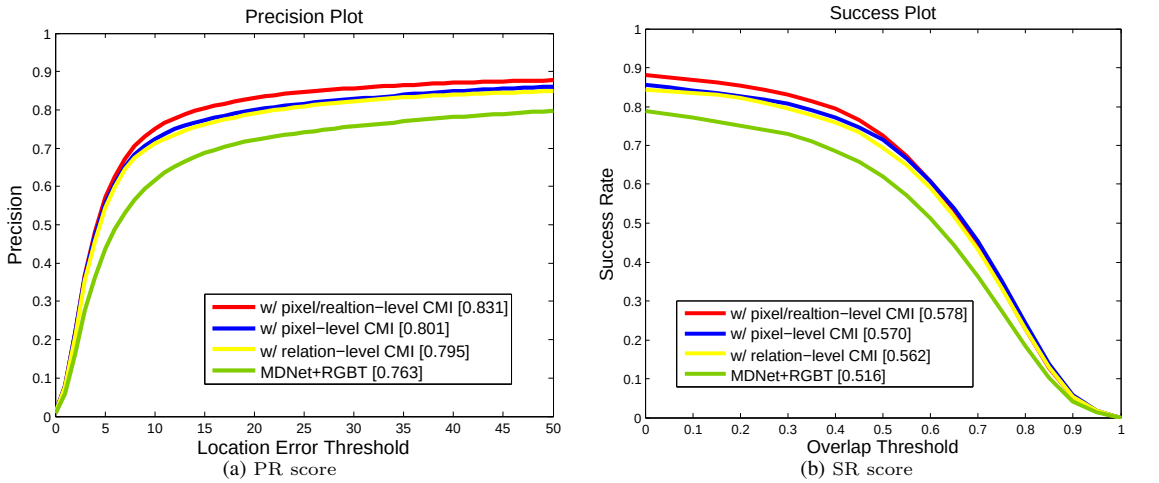
DAFNet [14], 3.4% over DAPNet [49], and 6.5% over MDNet [35]+RGBT in terms of the SR score. When compared with the top second CMPP method [38], a better tracking performance was achieved, for example, 93.9% *vs.* 92.6% in terms of the PR score and 74.1% *vs.* 73.8% in terms of the SR score. The above experimental results indicate that the proposed CMI method can effectively address the RGB-T tracking task by considering cross-modal interaction learning in the process of information fusion.

We further present the tracking performance comparisons between the proposed CMI method and other existing tracking methods on the VOT-RGBT2019 dataset, including TFNet [51], FANet [50], MANet++ [34], SGT [29], ATOM [8], SiamBAN [7], and ADRNet [46]. As shown in Table B3, the CMI obtains 0.625, 0.716, and 0.381 in the accuracy, robustness, and EAO metrics, respectively. Comparisons with previous tracking methods demonstrate that the proposed CMI method also substantially outperforms them, achieving improvement of 0.163 over TFNet [51], 0.153 over FANet [50], 0.116 over MANet++ [34], and 0.024 over DiMP [1] in terms of accuracy metric. In addition, we obtained the second and third ranks in the other two metrics, for example, 0.716 *vs.* 0.706 [7] for the robustness rate and 0.381 *vs.* 0.346 [27] for the EAO rate. Overall, this demonstrates that the proposed CMI method is effective in tracking video objects in RGBT video sequences.

Ablation study: To evaluate the effectiveness of the two CMI strategies, Fig. B3 shows the comparison results under various experimental settings on the RGBT234 dataset, including CMI with pixel- and relation-level interactions (i.e., "w/ pixel/relation-level CMI"), CMI with the pixel-level interaction (i.e., "w/ pixel-level CMI"), CMI with the relation-level interaction (i.e., "w/ relation-level CMI"), and a baseline method (i.e., MDNet+RGBT). In addition to adding the CMI module, our CMI and its variants have the same experimental settings as the baseline "MDNet+RGBT" method. When compared with the "MDNet+RGBT" method, our CMI with relation-level interaction achieves a better tracking performance, that is, 79.5% *vs* 76.3% in terms of PR score and 56.2% *vs* 51.6% in terms of SR score. When CMI network adopts the pixel-level interaction in the fusion process, the tracking performance achieves 80.1%/57.0% in PR/SR scores, which are also higher than the tracking results of "MDNet+RGBT". After simultaneously considering both the pixel- and relation-level interactions, our CMI achieves the best tracking performance, that is, 83.1%/57.8% in the PR/SR scores. The main reason for these improvements is that multi-level cross-modal interactions are considered during the information-fusion process, thus resulting in the effective capture of complementary characteristics between RGB and thermal images.

As shown in Fig. B4, we provide a qualitative comparison of our CMI against several state-of-the-art baselines for different RGB-T video sequences, including SGT [29], MDNet [35]+RGBT, SOWP [22]+RGBT, and C-COT [13]. Our CMI performs well in video sequences with low illumination and complex backgrounds, as illustrated in Fig. B4(a), and Fig. B4(b), respectively. This indicates that we can better predict the location of moving objects using our CMI, whereas other trackers lose the object of interest under extreme lighting conditions. In Fig. B4(b) and Fig. B4(d), tracking objects have a cluttered background and partial occlusion to some extent, and robust tracking results can also be obtained. This demonstrates that cross-modal fusion can help capture an effective representation of tracking objects and generate more precise predictions in challenging video sequences.

We present the runtime and tracking performance of our CMI against baseline methods, such as MDNet [36], MDNet+RGBT,



(a) PR score (b) SR score

Figure B3 Overall performance comparison with various experimental settings on the RGBT234 dataset.

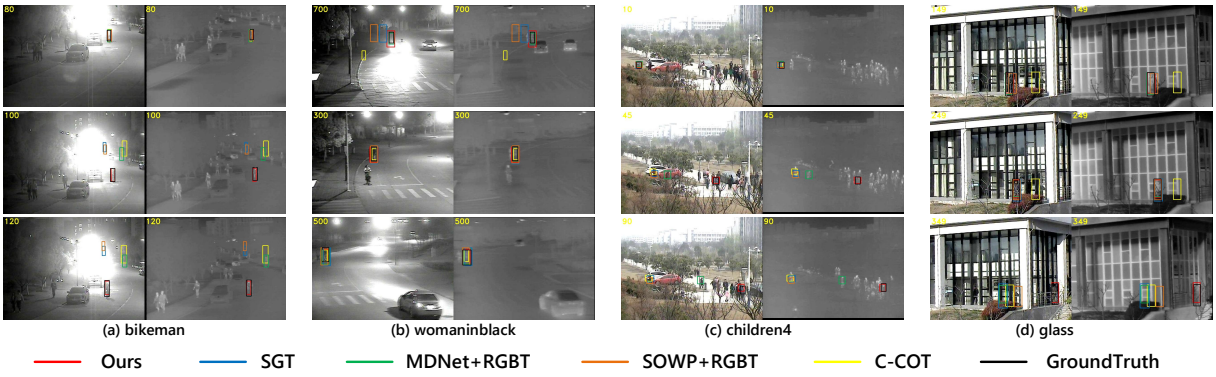


Figure B4 Qualitative comparison of our CMI against several state-of-the-art baselines on different RGB-T video sequences.

MANet [27], FANet [50], and CMPP [38]. For comparison, we used our CMI model on the platform of PyTorch with E5-2650@2.2 GHz CPU and NVIDIA GeForce 2080Ti GPU with an average tracking speed of 1.3 FPS. Note that the MDNet method [36] only employs RGB information to perform object tracking, whereas MDNet+RGBT simultaneously uses RGB and thermal information to address the RGB-T tracking problem. Compared with the baselines, we obtained a better tracking performance on the GTOT and RGBT234 datasets, except for MANet [27]. For example, we obtained 93.9%/74.1/% in PR/SR scores on the GTOT dataset and 83.1%/57.8% on the RGBT234 dataset with 1.3PFS, which are superior to MDNet [36] and MDNet+RGBT in both performance and runtime (i.e., MDNet+RGBT with 83.3%/67.6%/3.2FPS *vs.* ours 93.9%/74.1/%/1.6FPS on the GTOT dataset). The tracking performance of our CMI also outperformed that of FANet [50] and CMPP [38] with the same running speed. The above experimental results demonstrate that our tracking performance and speed can be improved through the use of the proposed CMI framework.

Table B4 Runtime(FPS) and performance(%)of our CMI against the baseline methods on GTOT and RGBT234 datasets.

Datasets	Metric	MDNet [36]	MDNet+RGBT	MANet [27]	FANet [50]	CMPP [38]	Ours
GTOT	PR	81.2	83.3	89.4	88.5	92.6	**93.9**
	SR	63.3	67.6	72.4	69.8	73.8	**74.1**
RGBT234	PR	71.0	76.3	77.7	76.4	82.3	**83.1**
	SR	49.0	51.6	53.9	53.2	57.5	**57.8**
	FPS	3.2	1.6	**1.1**	1.3	1.3	1.3

We further present the visual features and correlation maps of the CMI module to better understand the fusion process of the two heterogeneous modalities, including the inputs of RGB and thermal modalities, the correlations in the pixel and relation levels, and the outputs of the pixel-/relation-level CMI modules. As shown in Fig B5, the RGB and thermal modalities have different responses/representations for tracking objects. According to Eqn.(2) and Eqn.(3) in the revised manuscript, we can obtain the pixel-wise correlations of RGB and thermal modalities (i.e., $\mathcal{W}_{\mathcal{R}}^{\text{pixel}}$ and $\mathcal{W}_{\mathcal{T}}^{\text{pixel}}$) and the relation-level correlation of two modalities (i.e., $\mathcal{W}_{(\mathcal{R},\mathcal{T})}^{\text{relation}}$). We can obtain the output features by the pixel-level CMI and relation-level CMI strategies, which are aggregated and then fed into the subsequent binary classification network to boost the tracking performance. This indicates that the fused output features of the CMI can capture the appearance information of a specific object better than the input feature maps.

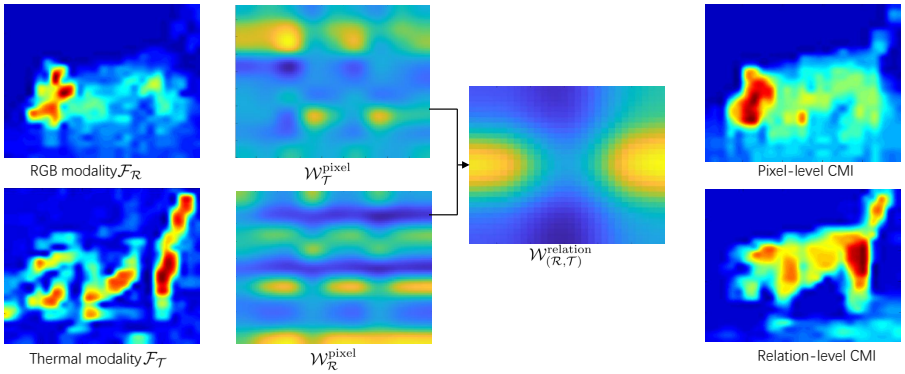


Figure B5 Visualization analysis for the CMI module, including the inputs of RGB and thermal modalities, the correlations in the pixel and relation levels, and the outputs of pixel-/relation-level CMI modules.

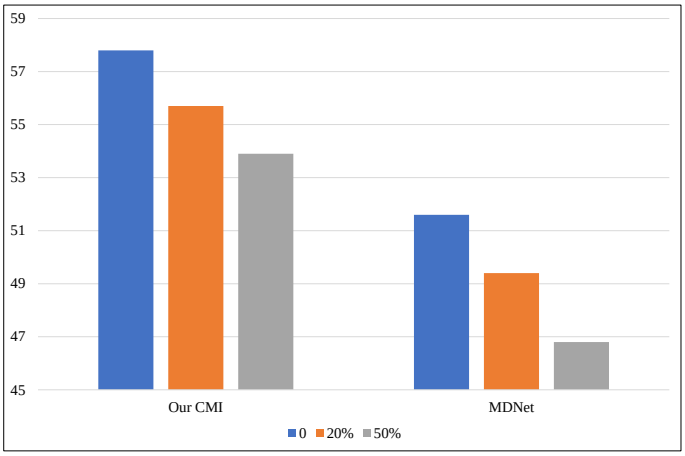


Figure B6 Performance comparisons (SR score) of CMI and MDNet under different degrees of modality loss on the RGBT234 dataset.

We conducted experiments to evaluate the robustness of the proposed CMI framework under the situation of modality loss, which is common in practice. Here, we randomly lost one of the modalities (either RGB or thermal information) for some frames in a video sequence, where the degrees of modality loss were set to 0, 20%, and 50%. As shown in Fig. B6, we report the performance comparisons (SR scores) of CMI and MDNet under different degrees of modality loss for the RGBT234 dataset. With a 20% degree of modality, the tracking results of CMI and MDNet were 55.7% and 49.4%, respectively, in terms of the SR score. When losing 50% of the modality information in a video sequence, the performance degradation of CMI and MDNet were 3.9% and 4.8%, respectively. In summary, these comparison results reveal that the adopted cross-modal interaction mechanisms are effective in addressing the RGBT tracking problem.

References

1 Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

2 David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550, 2010.

3 Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

4 K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C.L. Chen, and D. Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

5 L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

6 Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021.

7 Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020.

8 Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019.

9 Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017.

10 Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Convolutional features for correlation filter based visual tracking. In *IEEE International Conference on Computer Vision Workshop*, pages 621–629, 2015.

11 Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1561–1575, 2017.

12 Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. Adaptive color attributes for real-time visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014.

13 Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488, 2016.

14 Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *IEEE International Conference on Computer Vision Workshop*, pages 91–99, 2019.

15 R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

16 Ross B. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

17 K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

18 K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

19 João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2015.

20 João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge P. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *The 12th European Conference on Computer Vision*, volume 7575, pages 702–715, 2012.

21 Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time mdnet. In *Proceedings of European Conference Computer Vision*, Lecture Notes in Computer Science, pages 89–104, 2018.

22 Han-Ul Kim, Dae-Youn Lee, Jae-Young Sim, and Chang-Su Kim. Sowp: Spatially ordered and weighted patch descriptor for visual tracking. In *2015 IEEE International Conference on Computer Vision*, pages 3011–3019, 2015.

23 Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

24 Xiangyuan Lan, Mang Ye, Shengping Zhang, and Pong C Yuen. Robust collaborative discriminative learning for rgb-infrared tracking. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

26 Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

27 Chenglong Li, Andong Lu, Aihua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter RGBT tracking. In *IEEE International Conference on Computer Vision Workshops*, pages 2262–2270, 2019.

28 Chenglong Li, Xiang Sun, Xiao Wang, Lei Zhang, and Jin Tang. Grayscale-thermal object tracking via multitask laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):673–681, 2017.

29 Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of ACM Multimedia*, pages 1856–1864, 2017.

30 Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the ACM International Conference on Multimedia*, pages 1856–1864, 2017.

31 Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In *Proceedings of the European Conference on Computer Vision*, pages 808–823, 2018.

32 Yang Li, Zhan Xu, and Jianke Zhu. CFNN: correlation filter neural network for visual object tracking. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2222–2229, 2017.

33 Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *European conference on computer vision Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 254–265, 2014.

34 Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing*, 30:5613–5625, 2021.

35 Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

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

37 K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

38 Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for RGB-T tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7062–7071, 2020.

39 Naiyan Wang, Siyi Li, Abhinav Gupta, and Dit-Yan Yeung. Transferring rich feature hierarchies for robust visual tracking. *arXiv preprint arXiv:1501.04587*, 2015.

40 Naiyan Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 809–817, 2013.

41 Yi Wu, Erik Blasch, Genshe Chen, Li Bai, and Haibin Ling. Multiple source data fusion via sparse representation for robust visual tracking. In *14th International Conference on Information Fusion*, pages 1–8, 2011.

42 S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

43 Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10448–10457, 2021.

44 Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6728–6737, 2020.

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

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

47 Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4591–4600, 2019.

48 Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, pages 771–787. Springer, 2020.

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

50 Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust RGB-T tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2021.

51 Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo, and Liang Wang. Rgbt tracking by trident fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.