

Jointly Modeling Motion and Appearance Cues for Robust RGB-T Tracking

Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang[✉], Member, IEEE,

Huchuan Lu[✉], Senior Member, IEEE, and Xiaoyun Yang

Abstract—In this study, we propose a novel RGB-T tracking framework by jointly modeling both appearance and motion cues. First, to obtain a robust appearance model, we develop a novel late fusion method to infer the fusion weight maps of both RGB and thermal (T) modalities. The fusion weights are determined by using offline-trained global and local multimodal fusion networks, and then adopted to linearly combine the response maps of RGB and T modalities. Second, when the appearance cue is unreliable, we comprehensively take motion cues, i.e., target and camera motions, into account to make the tracker robust. We further propose a tracker switcher to switch the appearance and motion trackers flexibly. Numerous results on three recent RGB-T tracking datasets show that the proposed tracker performs significantly better than other state-of-the-art algorithms.

Index Terms—Visual tracking, RGB-T tracking, multimodal fusion.

I. INTRODUCTION

RGB-T tracking aims to integrate complementary visible (RGB) and thermal (T) infrared information to boost the tracking performance and make the tracker work in day and night [11]. First, the thermal infrared information is insensitive to illumination conditions and captures the target in extreme weather conditions, including night, fog, smog, to name a few. Second, the visible information is more discriminative

Manuscript received July 4, 2020; revised December 6, 2020 and February 2, 2021; accepted February 4, 2021. Date of publication February 26, 2021; date of current version March 4, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62022021, Grant 61806037, Grant 61872056, and Grant 61725202; in part by the Open Research Fund of the State Key Laboratory of Integrated Services Networks, Xidian University, under Grant ISN19-15; and in part by the Science and Technology Innovation Foundation of Dalian under Grant 2020JJ26GX036. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pavan Turaga. (*Corresponding author: Dong Wang*)

Pengyu Zhang and Dong Wang are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China, also with the Ningbo Institute, Dalian University of Technology, Ningbo 315016, China, and also with the Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: wdice@dlut.edu.cn).

Jie Zhao is with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China, and also with the Ningbo Institute, Dalian University of Technology, Ningbo 315016, China.

Chunjuan Bo is with the College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China.

Huchuan Lu is with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China, also with the Ningbo Institute, Dalian University of Technology, Ningbo 315016, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China.

Xiaoyun Yang is with Remark AI UK Ltd., London E14 9SH, U.K.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TIP.2021.3060862>, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3060862

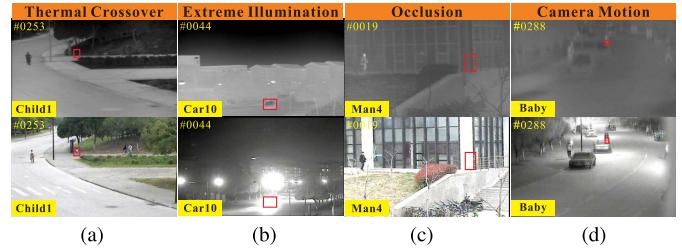


Fig. 1. Tracking with RGB-T modalities may suffer from thermal crossover, extreme illumination, occlusion, camera motion, and so on.

in foreground-background separation under normal circumstances and is more effective in separating two moving targets when thermal crossover occurs. Although many works have been done in recent years [5], [6], [12], [24], [68], effective fusion of RGB and T modalities and exploration of motion cues have big potential in designing a robust RGB-T tracker. Multimodal fusion is crucial to develop a robust appearance model in the RGB-T tracking task. As shown in Figure 1, information from each single modality is not always reliable, due to thermal crossover and extreme illumination. Existing methods mainly focus on information aggregation from different modalities. Li *et al.* [24] propose MANet to integrate both modality-shared and modality-specific information in an end-to-end manner. Zhang *et al.* [63] extend an RGB tracker to handle the RGB-T tracking task, and analyze different multimodal fusion types. Li *et al.* [26] focus on eliminating the modality discrepancy to exploit different properties between two modalities at the feature level. All aforementioned methods belong to early fusion, the strength of late fusion has not been explored in RGB-T tracking. In this work, we develop a novel late fusion method to generate the fusion weights of RGB and T modalities and use them to fuse the response maps for robust tracking. Motion information is also very important especially when the appearance cue is unreliable. Figure 1 shows that the target appearance dramatically changes when occlusion and camera motion occur. First, the appearance information is meaningless when the target is fully occluded (Figure 1 (c)). Second, camera moving with motion blur and out of search region makes the appearance model less effective (Figure 1 (d)). In aforementioned situations, we resort to some motion models for target prediction and camera motion compensation. Motion models have been exploited for traditional RGB tracking to deal with either target motion [21], [30] or camera motion [29], but few works have been done in the RGB-T tracking task. In this work, we attempt to

effectively model target and camera motions, acting as a vital supplement to the appearance model.

Motivated by the aforementioned discussions, this work attempts to jointly model the appearance and motion information for robust RGB-T tracking. Our contributions are summarized as follows.

- We propose a novel RGB-T tracking framework to take both appearance and motion information into account, which obtains a very robust performance. Specifically, both camera and object motions are embedded to assist appearance-based trackers. Besides, we propose an effective scheme to dynamically switch between appearance and motion cues.
- We propose a novel late fusion method (MFNet) to obtain both global and local weights for effective multimodal fusion, which results in a robust response map.

Numerous experimental results on three recent RGB-T tracking benchmarks, including GTOT, RGBT234, and VOT19-RGBT, show that our tracker performs significantly better than other competing algorithms. Our project page is available at: <https://github.com/zhang-pengyu/JMMAC>.

II. RELATED WORK

A. RGB-T Tracking

RGB-T tracking, as a branch of single target tracking, has drawn more attention in recent years [22], [23], [25], [26], [68]. Existing methods [5], [6], [24], [68] mainly focus on how to fuse the multimodal information for tracking. In [6], Conaire *et al.* propose an RGB-T framework to combine the feature representation from different modalities. After that, some methods improve tracking accuracy using sparse coding to mine multimodal information [26], [27], [33]. The JSR [33] method learns the joint sparse representation on different modalities and fuses the target likelihood using minimization operation. The LGMG [27] tracker constructs a multi-graph descriptor to suppress the background effects for RGB-T tracking. The CMR [26] algorithm utilizes a cross-modal manifold ranking algorithm to address the background clutter cases in RGB-T tracking. Recently, deep convolutional networks have been introduced into RGB-T tracking [24], [67], [68] and have significantly improved the tracking performance. FANet [67] learns both layer-wise and modality-wise feature weights to yield discriminative features for RGB-T tracking; while DAPNet [68] applies feature fusion and pruning process to achieve more robust feature representation. Furthermore, MANet [24] introduces a multi-adapter network to perform feature fusion in an end-to-end manner. For RGB-T tracking, it is important to exploit the information from both RGB and T modalities during the tracking process. In this work, we first develop a multimodal fusion network to conduct effective late fusion. Besides, we believe that motion cues are also very important for the RGB-T tracking task, and jointly model motion and appearance cues to improve the tracking accuracy.

B. Multimodal Fusion

Multimodal fusion attempts to integrate information from different modalities by employing the connection of multi-

modal data to obtain a more reliable classification or regression output [40], [43], [44], [60], [62], [65]. It usually can be categorized into two types: early fusion [3], [49] and late fusion [16], [51], [66]. Early fusion, by fusing low-level feature among modalities, aims to discover the complementary information of different modalities. Chaib *et al.* [3] propose a feature fusion method for very high resolution remote sensing scene classification. Shao *et al.* [49] conduct feature fusion for rotating machinery fault diagnosis via locality preserving projection. But the dimensionality of data expands multiply and thus leads to a high inference time. In contrast, late fusion combines the decisions (e.g., classification scores or tracking responses) to obtain a final result via the various fusion approaches, keeping independent models to give responses for each modality and maintaining flexibility [1]. Zheng *et al.* [66] propose a late fusion method at the score level to evaluate the quality of features. Terrades *et al.* [51] combine the result of classifiers via non-Bayesian probabilistic framework to improve the classification performance. Jain *et al.* [16] propose a score normalization approach for robust and fast late fusion. Recently, many researchers have applied early fusion to RGB-T tracking [24], [67], [68], while the effectiveness of late fusion has not been exploited. In this work, we first attempt to introduce late fusion to the RGB-T tracking task, thereby yielding better performance.

C. Tracking With Motion Cues

The motion cues (from target motion, camera motion or both) are also crucial but often ignored in designing a tracking framework. Target motion is widely used in previous tracking frameworks, such as Kalman filter [4], [18], [56] and particle filter [13], [21], [30], [42]. Comaniciu *et al.* [4] integrate Kalman filter and data association for target localization and representation. Weng *et al.* [56] utilize the adaptive Kalman filter to construct a motion model for tracking in complex situations (e.g., fast motion, occlusion and illumination variation). Kulikov and Kulikova [18] propose a continuous-discrete Extended Kalman filter method for radar tracking. Okuma *et al.* [42] propose an improved particle filter method, which combines mixture particle filter and Adaboost. Kwon *et al.* [21] develop a geometric particle filter technique based on matrix Lie groups, to overcome the limitation of the traditional deterministic optimization approach. As for camera motion, it is difficult to model camera motion since the camera parameter is unknown for inferring the 3D target location [29]. Furthermore, some datasets [22], [57], [58] are captured by stationary camera or with slight camera movement, where the importance of camera motion is underestimated. Existing works on tracking with drones predict camera movement since the camera is far away from the target and the depth variation is negligible [29]. Due to the powerful feature descriptions and learning algorithms, researchers have paid less attention to motion model when solving the tracking task [8], [41], [53]. However, in this work, we find that the consideration of target motion and camera motion could significantly improve the tracking performance in RGB-T tracking task.

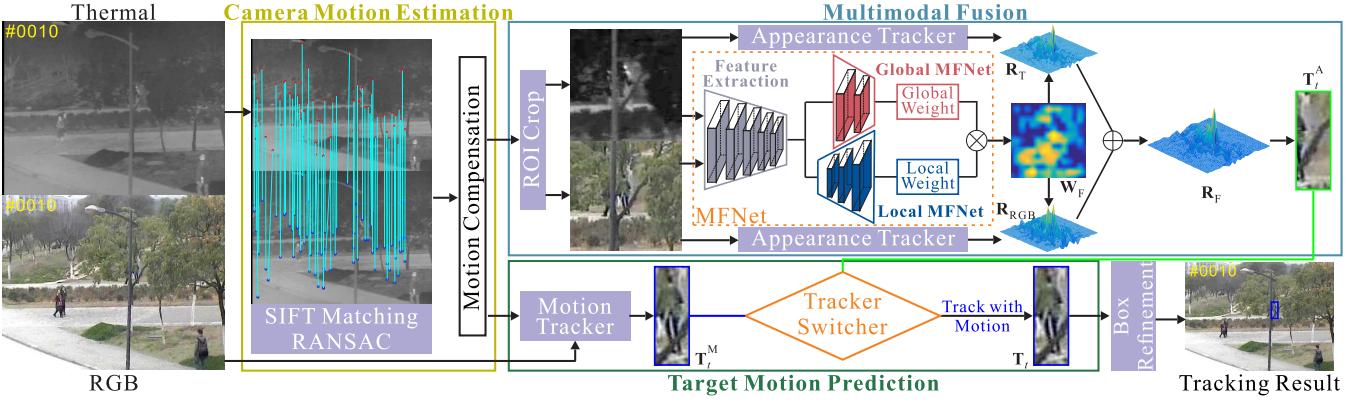


Fig. 2. JMMAC RGB-T tracking framework. We jointly model motion and appearance cues via two main components, i.e., multimodal fusion and motion mining. Multimodal fusion aims to fuse the appearance information in two modalities and improves the tracking accuracy by our MFNet. To the specific, we model motion cues using two schemes: target motion prediction and camera motion estimation. Target motion prediction predicts target position via motion information and determines which information is more reliable for tracking by tracker switcher. Camera motion estimation attempts to compensate for camera movement, stably providing effective search regions.

III. TRACKING FRAMEWORK VIA JOINTLY MODELING MOTION AND APPEARANCE CUES (JMMAC)

We propose a novel framework for robust RGB-T tracking by jointly modeling motion and appearance cues (JMMAC), which consists of two main components: multimodal fusion and motion mining. The JMMAC framework is shown in Figure 2. To be specific, we model motion cues using two schemes: target motion prediction and camera motion estimation. The tracking process can be easily summarized as follows. First, we apply camera motion compensation to deal with severe camera motion. Then, we exploit the late fusion method to aggregate tracking responses from different modalities via the proposed MFNet. We also maintain a motion tracker. When the appearance information is unreliable, our framework can dynamically select which cue is used for target location via target motion prediction module. Finally, we apply the bounding box refinement to adjust the final result for better scale estimation.

A. Multimodal Fusion Network for Robust Appearance Model

The ECO [8] method is chosen as our base tracker due to its effectiveness. Two ECO trackers are created for RGB and T modalities, respectively. Then, the corresponding response maps are obtained in each frame, denoted as $\mathbf{R}_{RGB} \in \mathbb{R}^{M \times N}$ and $\mathbf{R}_T \in \mathbb{R}^{M \times N}$ (the size of search regions is $M \times N$). In this work, we attempt to conduct a late fusion via a linear combination manner,

$$\mathbf{R}_F = \mathbf{W}_F \odot \mathbf{R}_{RGB} + (\mathbf{1} - \mathbf{W}_F) \odot \mathbf{R}_T, \quad (1)$$

where $\mathbf{W}_F \in \mathbb{R}^{M \times N}$ is a fusion weight whose elements are bounded from zero to one, and $\mathbf{1} \in \mathbb{R}^{M \times N}$ is a matrix whose elements are all one. \odot denotes the elementwise production operation. The target location can be determined based on the peak of the fused response map \mathbf{R}_F . The proposed MFNet, whose architecture is shown in Figure 2, aims to learn a precise pixel-wise fusion weight \mathbf{W}_F for RGB-T tracking from coarse to fine.

Note that our MFNet is effectively offline trained and directly applied for tracking without online fine-tuning.

MFNet. Our MFNet consists of a shared feature extractor and two subnetworks, namely global and local MFNet. As for the feature extractor, we use the truncated VGG-M network, which is pretrained on ImageNet without fine-tuning. Our MFNet takes the fixed size RGB and thermal patches $\mathbf{P}_{RGB} \in \mathbb{R}^{W \times H \times 3}$ and $\mathbf{P}_T \in \mathbb{R}^{W \times H \times 3}$ (the thermal patch is firstly expanded to a three-channel image) as inputs, and then extracts feature using the shared VGG-M network to obtain high-level features (Conv-5) \mathbf{F}_{RGB} and \mathbf{F}_T . W and H are the width and height of the image patch. After that, we concatenate them and send them to the subnetworks: (a) global MFNet, outputting a global weight $w_G \in \mathbb{R}^1$ to emphasize the contributions of different modalities; and (b) local MFNet, providing a pixel-level local weight $\mathbf{W}_L \in \mathbb{R}^{M \times N}$ to consider the distractors within each modality. The final weight $\mathbf{W}_F \in \mathbb{R}^{M \times N}$ can be obtained by

$$\mathbf{W}_F = w_G * \mathbf{W}_L, \quad (2)$$

where w_G and \mathbf{W}_L are constrained to $(0, 1)$ by using a Sigmoid layer. Figure 3 provides a visual example of the results obtained by our MFNet, indicating that the global weight w_G depicts the importance of different modalities and the local weight \mathbf{W}_L suppress the influence of distractors within each modality during the tracking process.

(1) Global MFNet. We first exploit the complementarity of RGB and T modalities by using a global MFNet to obtain the weight over the whole context. The output w_G of our global MFNet reflects the importance of each modality during the tracking process (see Figure 3). Our global MFNet contains two convolution layers, whose filter size is $3 \times 3 \times 256$ and $9 \times 9 \times 1$. Each layer is followed by rectified linear unit (ReLU) and local response normalization (LRN).

(2) Local MFNet. Though cross-modality divergence has been considered by our global MFNet, distractors in each individual modality is also harmful to robust tracking. Thus, we introduce a local MFNet to suppress the influence of distractors and achieve more accurate response maps. Our

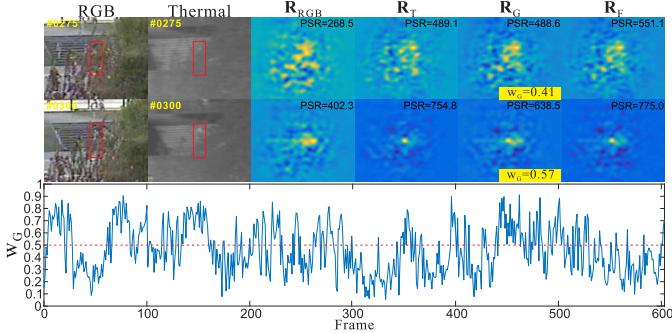


Fig. 3. Qualitative results of our MFNet. The global weight w_G dynamically varies to consider the contributions of two modalities during tracking process. The local weight \mathbf{W}_L provides a necessary supplement, resulting in a robust response map. The visualization of \mathbf{W}_L can be found in supplementary video.

local MFNet acts as an attention mechanism to obtain a finer weight map, and is constructed based on the U-Net architecture [48]. Specifically, our local MFNet consists of two deconv layers where the kernel size is $3 \times 3 \times 256$ and $3 \times 3 \times 1$ and a bilinear sampling layer which is adopted to resize the weight map to fit the size of response map. It is noted that each of deconv layers is followed by ReLU. In Figure 3, as a supplement to global MFNet (\mathbf{R}_G denotes the response only using global MFNet), local MFNet can suppress the distractors and obtains accurate response with high peak-to-sidelobe rate (PSR).

(3) Loss function. Similar to the classical correlation-filter (CF)-based methods [9], [15], we learn our MFNet by minimizing the squared Euclidean distance between the desired response \mathbf{Y} and the fused response map \mathbf{R}_F ,

$$\mathcal{L} = \|\mathbf{R}_F - \mathbf{Y}\|_2^2 \quad (3)$$

where the ground truth label \mathbf{Y} is a 2D Gaussian map whose peak denotes the target location.

B. Motion Modeling for Robust RGB-T Tracking

As presented in Section I and Figure 1, motion information is also very important for RGB-T tracking especially when the appearance information is unreliable due to complex variations (e.g., extreme illumination, low resolution, camera motion and occlusion). However, few existing RGB-T trackers have paid attention to mining motion cues. In this work, we explicitly divide motion information into target motion and camera motion, and design two different modules to handle them, namely, target motion prediction (TMP) and camera motion estimation (CME).

1) Target Motion Prediction: Appearance tracker cannot give accurate results when the appearance information is unreliable especially in low resolution and occlusion cases. In this situation, we design the target motion prediction scheme by introducing a motion tracker to predict target motion and proposing a tracker switcher to dynamic select which tracker is more reliable as final result.

Switcher. In our framework, we jointly model the appearance and motion cues by maintaining the appearance and motion trackers. We argue that in most cases, appearance is much more discriminative than motion information, while

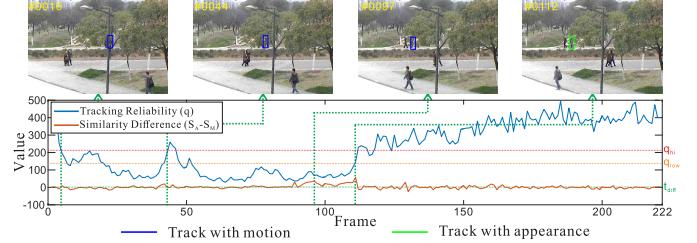


Fig. 4. Effectiveness of our TMP scheme. This figure indicates our TMP module works well in switching between appearance and motion trackers by combining MAX-PSR and template matching.

appearance can hardly help locate the target when the target is occluded or in low resolution. Thus, we design a simple yet efficient switching mechanism to determine which cue is more suitable for tracking, which simultaneously consider the reliability of response map and similarities between the target template and tracking results.

(1) MAX-PSR. First, we evaluate the reliability of the appearance tracker with a self-adaptive method presented in [36]. This method combines the PSR and maximum value of the response map \mathbf{R} . The reliability value q is defined as,

$$q = \text{PSR}(\mathbf{R}) \times \max(\mathbf{R}), \quad (4)$$

$$\text{PSR} = \frac{\max(\mathbf{R}) - \text{mean}(\mathbf{R})}{\text{var}(\mathbf{R})} \quad (5)$$

and $\max(\mathbf{R})$, $\text{mean}(\mathbf{R})$, $\text{var}(\mathbf{R})$ denote the maximum value, mean value and variance of the response map, respectively. The larger MAX-PSR is, the more reliable results we contain. However, as shown in Figure 4, the q value cannot always reflect the tracking reliability. Besides, online updating with noisy samples makes the response not sensitive to appearance variation. Thus, we develop the MAX-PSR method by considering the template similarity with template matching method to ensure a stable switching. The template matching scheme is not affected by noisy observations since it is without online update.

(2) Template Matching. We also compare the results of both appearance and motion trackers using the template matching method in RGB modality. First, we use \mathbf{T}_1 to denote the template image of target in the first frame. \mathbf{T}_t^A and \mathbf{T}_t^M are target regions obtained from appearance and motion trackers in the t -th frame. Then, we define two similarity scores s_A and s_M to evaluate the reliability of tracking results from appearance and motion trackers, respectively.

$$s_A = \mathcal{T}\mathcal{M}(\mathbf{T}_1, \mathbf{T}_t^A), \quad (6)$$

$$s_M = \mathcal{T}\mathcal{M}(\mathbf{T}_1, \mathbf{T}_t^M), \quad (7)$$

where $\mathcal{T}\mathcal{M}(\cdot)$ denotes the template matching function. To be specific, we use the deformable diversity similarity (DDIS) [50] method to conduct matching and calculate the similarity scores. The DDIS method measures the similarity between the target and template images based on the diversity of feature matches, and therefore is robust to complex deformation, background clutter and occlusion. More implementation details can be found in [50]. Once the MAX-PSR and template matching results are obtained, the tracker switcher

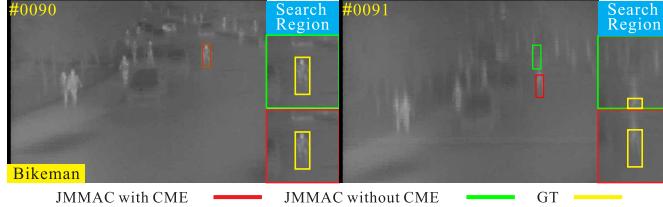


Fig. 5. Visual results of our tracker with and without CME.

determines which cue is employed for tracking as follows. If the appearance tracker is reliable enough (the MAX-PSR value q and template matching score S_A are both high). Furthermore, if the appearance tracker obtains a higher confidence than motion tracker (the MAX-PSR value q and template matching score S_A are relatively high, and the template matching score of appearance is larger than that of motion in a margin). Finally, if the appearance and motion tracker are all identified as an unreliable cue, we intuitively utilize the appearance cue to locate the target. The tracker switcher considers both offline information, i.e., the initial target template, and online response map, thus achieving good switch between the appearance and motion trackers. Figure 4 shows that our tracker switcher can select the meaningful information, thereby yielding a robust tracking result. The overall TMP scheme is summarized in Algorithm 1.

Algorithm 1 Target Motion Prediction (TMP)

Input: Fused response map \mathbf{R}_F , tracking results \mathbf{T}_t^A and \mathbf{T}_t^M , and the target template \mathbf{T}_1 .
Output: Final tracking result \mathbf{T}_t

- 1 Calculate the tracking reliability q via equation (4)
- 2 Obtain the similarities s_A and s_M via equations (6) and (7).
- 3 **if** ($q > q_{hi}$ and $s_A > s_{hi}$) or ($q > q_{low}$ and $s_A > s_{low}$ and $(s_A - s_M) > t_{diff}$) or $\max(s_A, s_M) < t_{disable}$ **then**
- 4 | $\mathbf{T}_t \leftarrow \mathbf{T}_t^A$ (using the appearance tracker);
- 5 | **else** $\mathbf{T}_t \leftarrow \mathbf{T}_t^M$ (using the motion tracker)
- 6 **end**

2) *Camera Motion Estimation:* Since RGB-T images are often captured by the high-altitude cameras being far from the targets, we assume that both target movement and depth variation of the target are very small. Thus, we directly model the camera motion in the 2D image plane rather than apply depth prediction. To be specific, we estimate camera motion by calculating the transformation matrix \mathbf{O} between the reference image $\mathbf{I}_r(x, y)$ and search image $\mathbf{I}_s(x, y)$ in thermal modality. The transformation matrix can be obtained by

$$(x', y') = \mathcal{T}(x, y; \mathbf{O}), \quad (8)$$

where (x', y') are coordinates of the key point in the search image corresponding to (x, y) in the reference image and $\mathcal{T}(\cdot)$ is the transformation function with the parameter matrix \mathbf{O} . The affine transformation with six parameters is adopted in default. First, key points of both reference and search images

are extracted with the scale-invariant feature transform (SIFT) method [35]. Then, the M-estimator sample consensus algorithm is used to match key points and exclude outliers. The obtained transformation matrix \mathbf{O} can compensate for the effect of camera motion. Figure 5 provides an example with large camera motion, indicating our CME scheme facilitates obtaining a stable search region.

C. Tracking With JMMAC

The overall tracking framework has been presented in Figure 2, some additional explanations are as follows.

Model Updating Scheme. Traditional CF-based methods update the filter every frame [9], [15] or fixed interval [8] with the fixed [15] or adaptive learning rate [10]. However, with such updating scheme, the filter may be degraded by the corrupted samples when the appearance information is unreliable. Thus, we skip the filter updating process when the motion tracker is applied to track the target. We update the motion tracker every frame to record the target trajectory generated by the appearance tracker. If the motion cues are utilized for tracking, the motion tracker predicts the target's position and the tracker is updated with the predicted result.

Motion Tracker. We apply Kalman filter tracker as our motion tracker. To make the paper self-contained, we detail the Kalman filter. We utilize Kalman filter to predict the target location by motion cues, i.e., position and velocity. In our method, we assume that target maintains constant velocity during sampling. In t -th frame, Kalman filter aims to estimate the target state $\mathbf{x}_t = (p_x, v_x, p_y, v_y)^T$ via a linear difference equation,

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_{t-1}, \quad (9)$$

where (p_x, p_y) denotes the coordinate of target center, (v_x, v_y) is the target velocity in both X-axis and Y-axis directions. \mathbf{A} is state transformation matrix which is defined as follow,

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (10)$$

\mathbf{w}_{t-1} denotes the process noise with normal probability distribution, i.e., $\mathbf{w}_{t-1} \sim N(0, \mathbf{Q})$, where \mathbf{Q} denotes the noise covariance matrix.

Two steps are included in Kalman filter tracking process, i.e., prediction and updating. In the prediction step, the target position is obtained via dynamic model expressed as,

$$\hat{\mathbf{z}}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t. \quad (11)$$

where \mathbf{x}_t and $\hat{\mathbf{z}}_t$ denote the target state and predicted measurement in the current frame, respectively. $\mathbf{H} \in \mathbb{R}^{2 \times 4}$ is measurement matrix which is defined as,

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (12)$$

and \mathbf{v}_{t-1} is the measurement noise and $\mathbf{v}_{t-1} \sim N(0, \mathbf{R})$. \mathbf{R} is the measurement covariance matrix. In the updating step, the target state is updated with the actual measurement \mathbf{z}_t .

More details can be found in [55]. In practice, the noise covariance \mathbf{Q} and measurement covariance \mathbf{R} are set to,

$$\mathbf{Q} = \begin{bmatrix} 25 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}, \quad (13)$$

$$\mathbf{R} = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}. \quad (14)$$

When the tracking result of Kalman filter is chosen as final result, Kalman filter is not updated since the actual measurement is unavailable. Otherwise, we consider the tracking result of appearance tracker as the actual measurement used for updating Kalman filter.

Bounding Box Refinement. The existing CF-based trackers estimate scale variation by sampling the search region among multiple resolutions and finding the corresponding scale with the maximum response. However, in this manner, the bounding box may not tightly capture the target since the ratio variation is ignored. In this work, we introduce a simple box refinement process followed by scale estimation with a real-time YOLOv2 [46] detector to alleviate this problem. The detector, which is pretrained on COCO [31] dataset, is efficiently applied in a small region surrounding the target without utilizing the category label. The refinement module is employed solely on visible modality, which avoids the data heterogeneity between the RGB and T images. Furthermore, since YOLOv2 obtains limited performance when detecting a small object and in extreme illumination, our refinement module is suspended in the aforementioned cases. The mean intensity of the whole image is utilized to detect extreme illumination. In the first frame, the intersection over union (IOU) between the initial target and predicted box, as well as its confidence, are used to judge the confidence of the YOLOv2 algorithm.

Implementation Details. We adopt a popular CF-based tracker, ECO [8], as our baseline. Both deep features (Conv1 and Conv5 of VGG-M for visible modality, Conv1, Conv4 and Conv5 of VGG-M for thermal modality) and handcraft features (HOG [7] and Color Name [52]) are used as feature representation. For training MFNet, we first train the global MFNet and freeze it, and then train the local MFNet. After that, we jointly fine-tune the overall MFNet. We random select training pairs (Two frames are included in a training pair. One is used for tracker initialization, another is used for learning MFNet.) within an interval of 5 frames, and then crop the ROI and obtain the training patches \mathbf{P}_{RGB} and \mathbf{P}_T . The resolutions of \mathbf{P}_{RGB} and \mathbf{P}_T are all set to 200×200 . The learning rate is set to $1e^{-5}$ when both global and local MFNet are trained and is set to $1e^{-7}$ in fine-tuning the overall MFNet, and the batch size is chosen as 8. The weight decay and momentum are set to 0.0005 and 0.9, respectively. We train our MFNet network using VOT19-RGBT when conducting evaluation on GTOT, and train it using GTOT when testing trackers on VOT19-RGBT and RGBT234. When drastic camera motion is detected, the multimodal fusion, target motion prediction and box refinement operations are suspended since the appearance information is unreliable. To further alleviate the computation

brought by CME and TMP, we apply a series of pre-processing operations. The CME pre-processing operation based on frame difference calculates the number of pixels with large variation between current and previous frames. Then, the frames with slight change are ignored to conduct CME module. Before applying template matching in TMP, the reliability value q is measured. If q is higher than 250, which indicates the appearance tracker is much more reliable for tracking, we suspend the template matching processing to boost the tracker. In target motion prediction module, the DDIS method extracts the deep features with the VGG-19 Network in default as described in [50]. The q_{hi} and sh_i are set to 210 and 15, q_{low} and s_{low} are set to 135 and 17, and t_{diff} is 3, respectively.

IV. EXPERIMENTS

We evaluate our tracker in comparison with other competing ones using three RGB-T tracking datasets (GTOT [22], RGBT234 [23], and VOT19-RGBT [17]). Our tracker (referred as **JMMAC**) is implemented by MATLAB 2015b, Intel i9 CPU with 64G RAM and a RTX2080Ti GPU with 11G memory. *Both training and testing codes will be publicly available.*

Datasets and Metrics. The GTOT dataset [22], constructed in 2016, contains 50 grayscale-thermal sequences annotated with seven challenging attributes, including occlusion (OCC), large-scale variation (LSV), fast motion (FM), low illumination (LI), thermal crossover (TC), small object (SO), and deformation (DEF). The RGBT234 dataset [23], proposed in 2019, is the largest RGB-T tracking benchmark at present. This dataset includes 234 sequences with more than 234,000 frames and extends the number of attributes to 12. These attributes include 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). The comparisons follow the one pass evaluation (OPE) rule with Success Rate (SR) and Precision Rate (PR), which are widely used in [57], [58]. Given that two modalities exist in the RGB-T tracking task, the SR and PR values are used to measure the maximum Intersection over Union (IoU) and the minimum center location error between two modalities frame by frame (denoted as MSR and MPR [22], [23]). The VOT19-RGBT [17] challenge selects 60 representative video clips from RGBT234 [23] and adopts the Expected Average Overlap (EAO) rule (along with tracking accuracy (A) and robustness (R)) to emphasize the short-term evaluation. Besides, the EAO values are calculated based on the trackers' results and the ground truths of the thermal modality (we refer the readers to [17] for more details).

Compared Algorithms. By using GTOT [22] and RGBT234 [23], we compare our tracker with nine competing RGB-T methods, including MANet [24], FANet [67], ECO+RGBT, TODA [61], DAPNet [68], SiamFC+RGBT, SGT [25], CMR [26] and mfDiMP [63].¹ All these trackers have achieved

¹ECO+RGBT and SiamFC+RGBT are our implemented trackers, which improve the traditional ECO [8] and SiamFC [2] ones by concatenating RGB and thermal features as input features.

TABLE I

ATTRIBUTE-BASED COMPARISON WITH NINE STATE-OF-THE-ART TRACKERS ON THE GTOT DATASET. MAXIMUM SUCCESS RATE AND MAXIMUM PRECISION RATE (MSR\MPR %) ARE USED FOR EVALUATION. IT IS INDICATED THAT JMMAC OUTPERFORMS ALL THE COMPETITORS WITH A LARGE MARGIN

	OCC	LSV	FM	LI	TC	SO	DEF	ALL
MANet	69.6/ 88.2	70.1/87.6	69.9/87.6	71.2/89.0	71.0/89.0	70.8/89.7	71.5/90.1	72.4/89.4
FANet	70.3 /86.4	69.2/84.0	68.4/83.2	70.2/89.9	70.8/86.7	70.8/88.1	71.8/89.1	72.8/89.1
ECO+RGBT	66.4/81.1	69.3/83.7	70.0/83.5	70.7/85.4	69.6/84.2	69.3/85.0	69.0/85.0	67.7/82.1
TODA	63.5/84.6	64.4/84.8	64.2/84.8	66.3/85.1	66.0/85.3	66.3/86.7	67.6/86.9	67.7/84.3
DAPNet	67.4/87.3	66.1/86.0	65.3/85.2	67.7/86.9	68.0/87.5	68.2/88.6	69.6/89.1	70.7/88.2
SiamFC+RGBT	59.3/74.7	62.2/77.7	61.1/75.8	60.5/74.7	61.1/76.0	60.3/76.1	60.1/75.2	60.6/74.1
SGT	56.7/81.0	55.7/82.6	55.7/82.0	59.0/84.3	59.6/84.4	60.0/85.7	62.1/86.7	62.8/85.1
CMR	62.6/82.5	64.7/83.9	64.7/83.8	65.8/85.5	64.9/84.4	64.2/84.8	64.4/84.8	64.3/82.7
mfDiMP	45.4/54.8	47.8/57.2	49.0/57.4	47.8/56.0	48.6/58.6	48.4/59.8	48.3/59.8	49.0/59.4
JMMAC	68.4/84.0	71.5/87.4	72.4/87.9	73.9/90.3	73.0/89.9	73.2/90.7	73.6/91.7	73.2/90.1

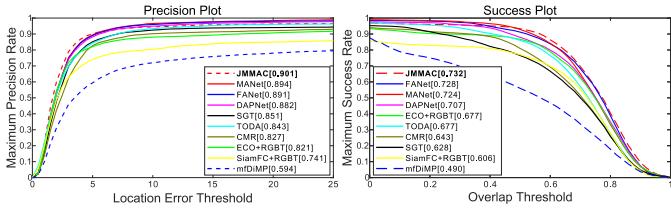


Fig. 6. Performance evaluation on GTOT in terms of success and precision plots.

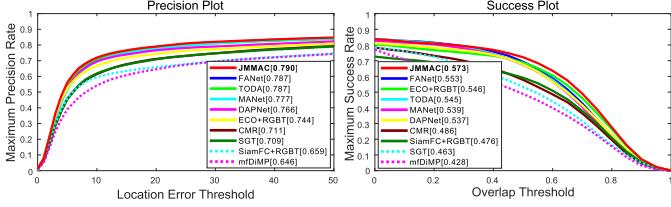


Fig. 7. Performance evaluation on RGBT234 in terms of success and precision plots.

top-ranked performance or serve as baselines in both GTOT and RGBT234 benchmarks and they can be categorized into deep-learning-based (MANet, SiamFC+RGBT, TODA, ECO+RGBT, DAPNet, mfDiMP and FANet), CF-based (ECO+RGBT), graph-learning-based (SGT) and sparse-coding-based (CMR) ones. In addition, we compare our JMMAC method with eight recent trackers in the official VOT2019 [17] challenge report.

A. Quantitative Evaluation

GTOT [22]. First, we compare our tracker with other methods using the GTOT dataset and report both success and precision plots in Figure 6. We can see that our JMMAC method obtains the best performance with 73.2% and 90.1% in success and precision scores, respectively. Compared with the most recent tracker (also the second best one), MANet, our algorithm achieves 0.8% improvement in success and 0.7% in precision. As shown in Table I, the attribute-based comparison also shows the capability in handling occlusion (OCC), large scale variation (LSV), fast motion (FM), low illumination (LI), thermal crossover (TC), small object (SO) and deformation (DEF).

RGBT234 [23]. Second, we evaluate JMMAC method using the large-scale RGBT234 dataset and report the related results in Figure 7. Overall, our tracker is superior to all the compared

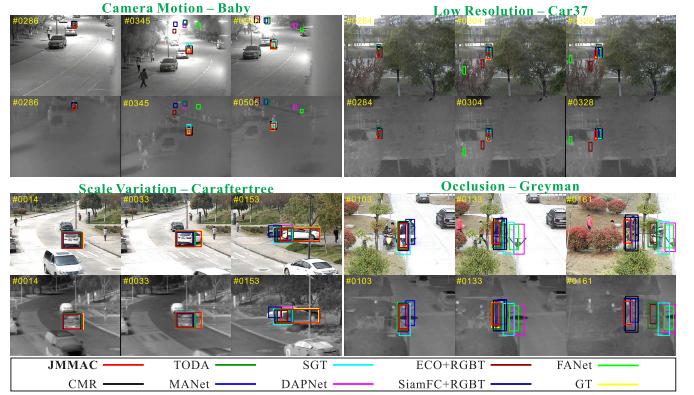


Fig. 8. Representative visual results of our JMMAC and other state-of-the-art trackers on the VOT-RGBT dataset.

algorithms. Table II summarizes the JMMAC's performance in handling different challenging factors. Our tracker works very well in dealing with occlusion, low illumination, deformation, scale variation, motion blur, and camera moving. Those challenges result in unreliable appearance information and therefore renders the tracker easy to drift. An important reason is that our method develops an effective multimodal fusion network to learn a robust appearance model. Additionally, our camera motion compensation and target motion prediction schemes further alleviate the effects of unreliable appearance information.

VOT19-RGBT [17]. Finally, we test our tracker using the VOT19-RGBT dataset and report EAO, A and R values in Table III. The results of other competing methods are obtained from the official VOT2019 [17] challenge report. Our JMMAC tracker performs the best for all three metrics with significant performance superiority. Compared with the second-rank tracker (SiamDW_T), We achieve 26.8% relative improvement in EAO with more accurate results and less failure times.

B. Qualitative Evaluation

Figure 8 illustrates some qualitative results from VOT19-RGBT to compare our JMMAC and with other trackers in handling several challenging cases, such as camera motion (*Baby*), low resolution (*Car37*), scale variation (*Caraftertree*), and occlusion (*Greyman*). Our JMMAC tracker, which utilizes

TABLE II

ATTRIBUTE-BASED EVALUATION WITH NINE STATE-OF-THE-ART TRACKERS ON THE RGBT234 DATASET. MAXIMUM SUCCESS RATE AND MAXIMUM PRECISION RATE (MSR/MPR) ARE USED FOR EVALUATION. THE TOP PERFORMANCE IS MARKED IN red FONT

	MANet	FANet	ECO+RGBT	TODA	DAPNet	SiamFC+RGBT	SGT	CMR	mfDiMP	JMMAC
NO	64.6/88.7	65.7/88.2	65.7/88.5	64.6/89.3	64.4/90.0	61.2/81.6	55.9/86.8	61.6/89.5	50.5/73.6	69.4/93.2
PO	56.6/81.6	60.2/ 86.6	58.8/80.3	57.2/82.7	57.4/82.1	49.4/69.3	49.0/74.8	53.6/77.7	48.3/71.9	61.1/84.1
HO	46.5/68.9	45.8/66.5	45.4/62.2	47.4/ 69.8	45.7/66.0	39.7/55.6	39.2/59.9	37.7/56.3	33.5/53.0	48.3/67.7
LI	51.3/76.9	54.8/80.3	58.8/82.9	55.3/80.3	53.0/77.5	47.7/67.4	44.4/68.7	49.8/74.2	43.8/64.7	58.8/84.0
LR	51.5/75.7	53.2/79.5	48.6/71.5	52.2/78.4	51.0/75.0	42.5/62.1	48.0/75.6	42.0/68.7	37.5/62.8	51.7/77.1
TC	54.3/75.4	54.9/76.6	57.6/77.9	50.7/74.0	54.3/76.8	44.4/60.8	45.3/72.7	44.3/67.5	44.9/69.5	52.6/74.9
DEF	52.4/72.0	52.6/72.2	49.8/66.3	51.6/74.3	51.8/71.7	46.1/62.4	46.6/67.7	47.3/66.7	42.7/60.7	52.9/70.6
FM	44.9/69.4	43.6/68.1	43.9/63.8	48.0/75.3	44.3/67.0	38.6/58.0	39.0/66.6	38.4/61.3	35.4/55.7	41.7/61.0
SV	54.2/77.7	56.3/78.5	57.4/76.2	55.4/79.2	54.2/78.0	48.9/67.1	43.3/69.3	49.3/71.0	47.0/68.3	61.6/83.7
MB	51.6/72.6	50.3/70.0	49.9/67.0	50.1/70.7	46.7/65.3	39.2/52.9	42.0/62.2	42.7/60.0	42.0/61.9	54.9/75.1
CM	50.8/71.9	52.3/72.4	51.3/68.9	49.3/69.8	47.4/66.8	43.6/59.6	43.8/64.8	44.7/62.9	41.9/61.4	55.6/76.2
BC	48.6/73.9	50.2/75.7	47.3/67.9	51.3/77.1	48.4/71.7	41.0/57.9	40.4/63.9	39.8/63.1	33.4/54.7	48.5/68.7
ALL	53.9/77.7	55.3/78.7	54.6/74.4	54.5/78.7	53.7/76.6	47.6/65.9	46.3/70.9	48.6/71.1	42.8/64.6	57.3/79.0

TABLE III

COMPARISON RESULTS ON VOT19-RGBT. JMMAC OUTPERFORMS ALL THE COMPETITORS BY A LARGE MARGIN. THE TOP THREE RESULTS ARE IN red, blue AND green FONTS

Trackers	GESBTT	CISRDCF	MPAT	MANet	FSRPN	mfDiMP	SiamDW_T	JMMAC
A (\uparrow)	0.6163	0.5215	0.5723	0.5823	0.6362	0.6019	0.6158	0.6597
R (\uparrow)	0.6350	0.6904	0.7242	0.7010	0.7069	0.8036	0.7839	0.8235
EAO (\uparrow)	0.2896	0.2923	0.3180	0.3463	0.3553	0.3879	0.3925	0.4978

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT FUSION METHODS ON THE IMAGE FUSION TASK USING THE DATASET PRESENTED IN [37]. OUR MFNET ACHIEVES THE COMPETITIVE PERFORMANCE WITH REAL-TIME SPEED

Method	EN (\uparrow)	MI (\uparrow)	Qabf (\uparrow)	FMI _w (\uparrow)	Nabf (\downarrow)	SSIM (\uparrow)	MS_SSIM (\uparrow)	FPS (\uparrow)
DenseFuse [28]	6.841	13.683	0.448	0.432	0.081	0.710	0.932	1.8
FusionGAN [37]	6.572	13.144	0.234	0.392	0.078	0.631	0.748	3.6
CBF [20]	6.907	13.815	0.414	0.319	0.331	0.563	0.672	0.047
WLS [38]	6.821	13.642	0.490	0.377	0.223	0.691	0.930	0.85
JSR [64]	6.575	13.149	0.376	0.223	0.222	0.601	0.867	0.0033
JSRSD [32]	6.884	13.767	0.343	0.199	0.319	0.539	0.788	0.0027
CSR [34]	6.433	12.866	0.531	0.388	0.021	0.723	0.906	0.0074
MFNet (Ours)	6.912	13.825	0.425	0.428	0.078	0.714	0.895	26.6

both appearance and motion cues, can capture the tracked object accurately in camera motion, low resolution and occlusion cases. And, JMMAC with box refinement module can obtain a more tight bounding box in sequence ‘Caraftertree’, where the object has large size variation.

C. Ablation Analysis

Effectiveness of Different Components. To provide a thorough analysis of each component, we compare several variants of our JMMAC tracker, including: (1) JMMAC(B)-RGB: the baseline method with only RGB modality; (2) JMMAC(B)-T: the baseline method with only thermal modality; (3) JMMAC(B)+MF-G: JMMAC(B) only with Global MFNet; (4) JMMAC(B)+MF-L: JMMAC(B) only with Local MFNet; (5) JMMAC(B)+MF: JMMAC(B) with MFNet (i.e., the combination of both global and local MFNets); (6) JMMAC(B)+MF+CME: JMMAC(B) with multimodal fusion (MF) and camera motion estimation (CME) modules; (7) JMMAC(B)+MF+CME+TMP: JMMAC(B) with multimodal fusion (MF), camera motion estimation (CME) and target motion prediction (TMP) modules; (8) JMMAC: our final model, i.e., JMMAC(B)+MF+CME+TMP+BR. ‘BR’ means that the tracker exploits the detection scheme to refine the output bounding boxes if necessary. The detailed comparison results on all three datasets are reported in Table V.

Each component substantially contributes to the final JMMAC tracker and gradually improves the tracking performance. Both global and local fusion networks can effectively improve the baseline. Compared with the data shown in Table III, our tracker performs better than other competing algorithms when used with multimodal fusion only (see JMMAC(B)+MF vs SiamDW_T), which shows the effectiveness of the proposed MFNet. Since the sequences are captured by still camera in GTOT, CME module does not influence the tracking result.

Effectiveness of local MFNet. Here, we detail the effectiveness of our local MFNet on suppressing distractors. In general, distractor denotes the candidate or background region whose appearance or semantic information is similar to the tracked target in any modality. Our local MFNet functions as an attention module by comparing which modality is more reliable/informative in modality-level and pixel-level and suppresses distractors that may lead to multi-peak responses. As shown in Figure 9, two women in different colors walk side by side in sequence ‘Woman89’. The RGB tracker can recognize the target and yields a stable response, while the T tracker provides a multi-peak response. MFNet provides the modality confidence by comparing which modality carries more information and is more reliable for tracking, thereby giving a larger weight on the region where the woman in black is in the RGB modality and suppressing the response

TABLE V
EFFECTIVENESS OF EACH COMPONENT FOR JMMAC ON THE VOT19-RGBT, GTOT AND RGBT234 DATASETS. THE TOP THREE RESULTS ARE IN RED, BLUE AND GREEN FONTS

Trackers	VOT19-RGBT			GTOT		RGBT234	
	EAO (\uparrow)	A (\uparrow)	R (\uparrow)	MSR (\uparrow)	MPR (\uparrow)	MSR (\uparrow)	MPR (\uparrow)
JMMAC(B)-RGB	0.3207	0.5909	0.6987	62.1	75.3	52.8	72.2
JMMAC(B)-T	0.3862	0.6452	0.7604	65.4	74.2	51.6	71.1
JMMAC(B)+MF-G	0.4102 (+2.50%)	0.6502	0.7835	67.5 (+2.1%)	81.5 (+6.2%)	54.3 (+1.5%)	74.6 (+2.4%)
JMMAC(B)+MF-L	0.4073 (+2.11%)	0.6387	0.7835	67.7 (+2.3%)	79.8 (+4.5%)	55.4 (+2.6%)	75.4 (+3.2%)
JMMAC(B)+MF	0.4116 (+2.54%)	0.6465	0.7835	70.5 (+5.1%)	85.1 (+9.8%)	55.6 (+2.8%)	75.4 (+3.2%)
JMMAC(B)+MF+CME	0.4198 (+0.82%)	0.6505	0.7953	70.5 (+0.0%)	85.1 (+0.0%)	56.0 (+0.4%)	76.4 (+1.0%)
JMMAC(B)+MF+CME+TMP	0.4598 (+4.00%)	0.6497	0.8073	70.9 (+0.4%)	85.5 (+0.4%)	56.4 (+0.4%)	77.1 (+0.7%)
JMMAC	0.4978 (+3.80%)	0.6597	0.8235	73.2 (+2.3%)	90.1 (+4.6%)	57.3 (+0.9%)	79.0 (+1.9%)

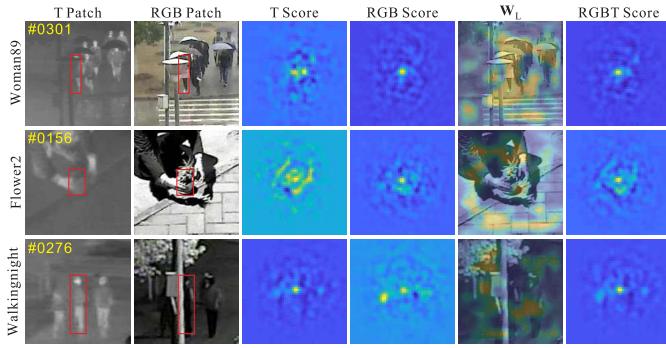


Fig. 9. Effectiveness of our Local MFNet in handling distractors.

of the distractor. In sequence ‘Flower2’, the vase is salient in the RGB modality while the vase’s temperature is similar to the background. The RGB tracker can obtain a more reliable score than the T tracker. W_L with higher weights on the target region suppresses the T score and obtains a reliable result. In sequence ‘Walkingnight’, since the target is in the dark, its RGB cue is less informative than the corresponding thermal cue and the RGB tracker drifts to the distractor. W_L gives lower weights on the target region in the RGB modality, thereby suppressing the confidence of the RGB modality and obtaining a correct tracking result.

MFNet analysis on image fusion. We find that the goals of image fusion and multimodal fusion in tracking are similar. In tracking, we assume that the more possible the target is, the more salient information the modality contains, while the image fusion aims to combine the visible and thermal images while preserving both thermal radiation and detailed texture information. To this end, our learned MFNet can be also generalized to deal with the image fusion task (even without fine-tuning on related datasets). As described in Section III-A, we send the RGB-T images to MFNet and obtain the fused image via $\mathbf{I}_F = \mathbf{W}_F \times \mathbf{I}_{RGB} + (1 - \mathbf{W}_F) \times \mathbf{I}_T$. We compare our MFNet with other algorithms using the dataset presented in [37], which includes 41 pairs of testing images collected from TNO and INO datasets. The TNO dataset² contains image sequences registered with different multiband camera systems, recording multi-spectral nighttime imagery of different military relevant scenarios. The INO dataset,³ constructed

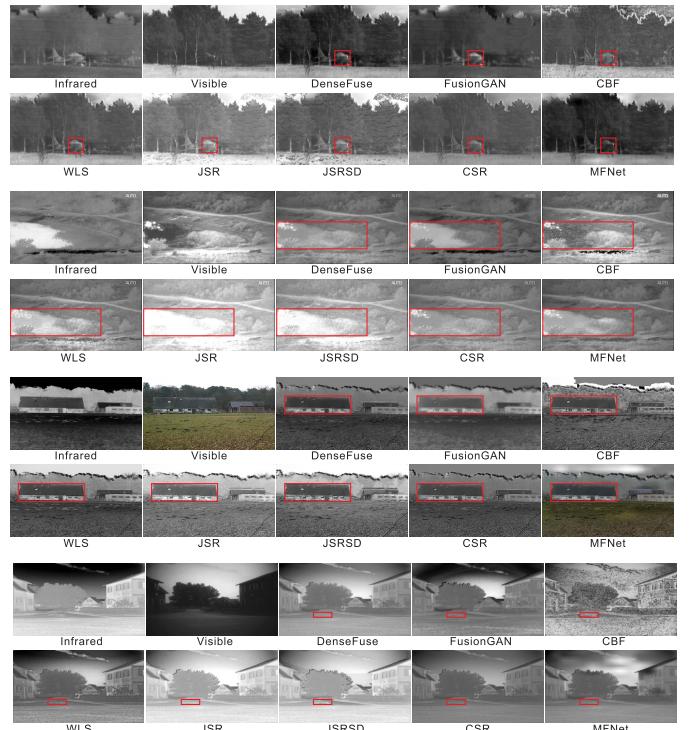


Fig. 10. Visualization results on image fusion task.

by the National Optics Institute of Canada, consists of several image pairs and sequences captured in different conditions.

In this paper, we choose eight common metrics for evaluation, including entropy (EN) [47], mutual information (MI) [45], edge information (Qabf) [59], feature mutual information (FMI) [14], fusion artifacts (Nabf) [19], structural similarity index measure (SSIM) [54], multi-scale SSIM (MS_SSIM) [39], and inference speed (frame per second, FPS). The comparison results in Table IV indicate that our MFNet achieves very competitive performance with real-time performance. Our MFNet performs consistently better than FusionGAN (the second fast method) with speeds seven time faster. Furthermore, we show 8 pairs of fusion results to validate that our MFNet can also achieve very competitive results on the image fusion task. The testing images are from previous work [37]. As shown in Figure 10, MFNet can fuse complementary information from both modalities while retaining more detail information in the visible modality.

²https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029

³<https://www.ino.ca/en/video-analytics-dataset/>

TABLE VI

COMPARISONS OF DIFFERENT CAMERA MOTION MODELS. HERE, WE ADOPT THE JMMAC(B)+MF VARIANT AS THE BASELINE

Motion models	EAQ (\uparrow)	A (\uparrow)	R (\uparrow)
JMMAC(B)+MF	0.4116	0.6465	0.7835
JMMAC(B)+MF+Translation	0.3825	0.6527	0.7529
JMMAC(B)+MF+Similarity	0.4163	0.6476	0.8073
JMMAC(B)+MF+Affine	0.4198	0.6505	0.7953
JMMAC(B)+MF+Projective	0.4136	0.6516	0.7835

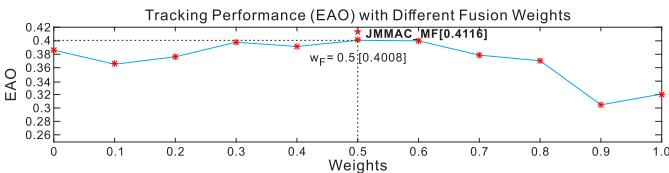


Fig. 11. Tracking performance with different fusion weights w in VOT19-RGBT.

Analysis of fusion methods. We also implement some variants with earlier or late fusion manners, and find that the proposed MFNet performs the best. We apply 5 other fusion methods to validate the strength of MFNet, which contains both early fusion and late fusion methods. (1) Merge: layer-wise add the features in RGB and thermal modalities. (2) Concatenate: layer-wise concatenate the features in RGB and thermal modalities. (3) Concatenate+PCA: We apply feature concatenation followed by Principal Component Analysis (PCA) operation. (4) Intensity-based fusion: we assume that the temperature of object is constant and we calculate the intensity of target i_1 in the first frame. we add a penalty to the response expressed as,

$$\mathbf{R}_F = \frac{1}{2} \mathbf{P} \times (\mathbf{R}_{RGB} + \mathbf{R}_T) \quad (15)$$

$$s.t. \quad \mathbf{P}(i, j) = \min\left(\frac{i_t(i, j)}{i_1}, \frac{i_1}{i_t(i, j)}\right) \quad (16)$$

where (i, j) denotes the location in response map and i_t denotes the intensity of target in the t -th frame. (5) Tracking-quality-based fusion: we fuse the responses with the guidance of tracking quality described in [36] and the final response can be expressed as,

$$\mathbf{R}_F = \frac{q_{RGB}}{q_T + q_{RGB}} \times \mathbf{R}_{RGB} + \frac{q_T}{q_T + q_{RGB}} \times \mathbf{R}_T \quad (17)$$

where q_{RGB} and q_T are tracking quality of RGB and thermal modalities calculated by Equation (5) in our submission, (6) MFNet: our proposed approach. Also, we add the comparison between the trackers solely with RGB and thermal information. From Table VII, all fused methods can improve tracking performance except for merge and tracking-quality-based fusion. This may be caused by the discrepancy of data from different modalities and the uncertainty of tracking quality. MFNet outperforms all other fusion methods in a large margin. Late fusion, which can extract features in different layers and fuse them in various scales, is more flexible than early fusion, that fusion operation only can be applied to feature in the same size. Furthermore, We fuse the responses via a weighted linear combination, i.e., $\mathbf{R}_F = w \times \mathbf{R}_{RGB} + (1 - w) \times \mathbf{R}_T$. In Figure 11, we enumerate fusion weight w from 0 to 1 with

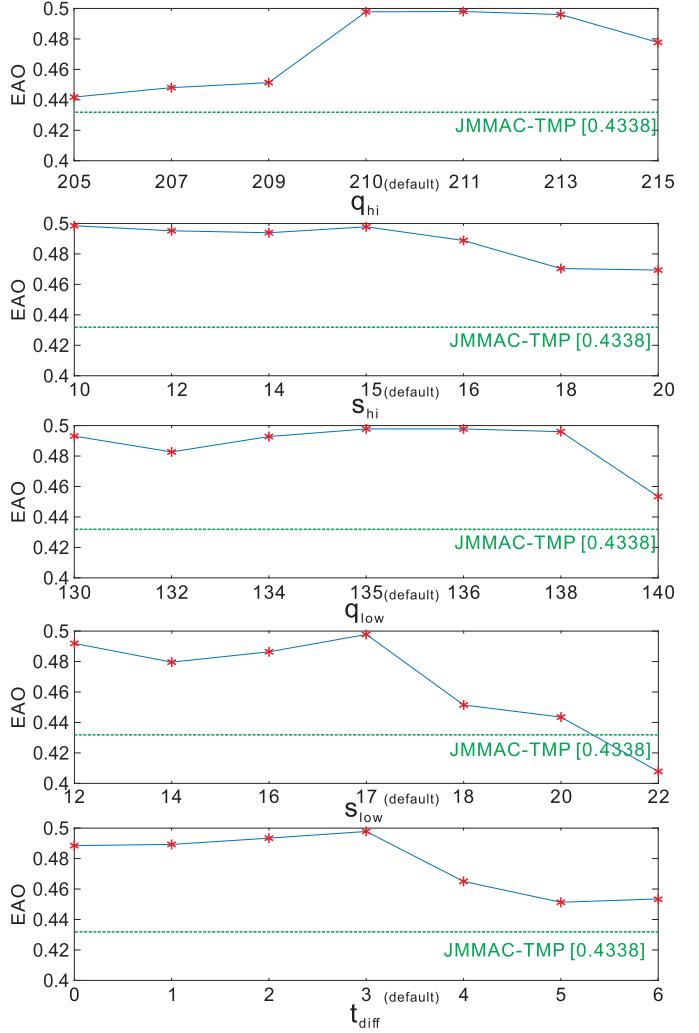


Fig. 12. Parameter robustness analysis. We enumerate the parameters with the interval 2 (especially, for t_{diff} , we set the interval to 1) individually. Our tracker switcher is robust against parameter perturbation and works well in a large range.

an interval 0.1 and report the tracking performance (EAQ) on VOT19-RGBT.

Parameter Robustness Analysis of JMMAC. Since both online information (quality of response) and offline information (template in the initial frame) are considered to jointly determine which cue is used for tracking, we argue that our target motion prediction (TMP) module is very robust against parameter perturbation and the parameters are not over-fitting to specific dataset. To fully validate this, we enumerate the all the parameters with interval 2 (for t_{diff} , we set the interval to 1) and report their EAQs in VOT19-RGBT. As shown in Figure 12, compared with the tracker without TMP module (JMMAC-TMP for short), JMMAC with TMP achieves performance promotion with a large parameter range.

Different Camera Motion Models. We also test the effects of different camera motion models, including translation, similarity, affine, and projective transformations between frames. The comparison results are summarized in Table VI. All camera motion models, except for translation transformation, can improve the tracking performance. Among those models,

TABLE VII

FUSION METHODS ANALYSIS ON VOT19-RGBT. BOTH EARLY FUSION AND LATE FUSION METHODS ARE INCLUDED FOR EVALUATION. COMPARED WITH DIFFERENT TYPE OF FUSION METHODS, OUR MFNET OBTAINS THE BEST PERFORMANCE

Fusion type	Fusion method	Available modality	EAO (\uparrow)	A (\uparrow)	R (\uparrow)
Single modality	-	RGB	0.3207	0.5909	0.6987
	-	Thermal	0.3862	0.6452	0.7604
Early fusion	Merge	RGB & Thermal	0.3734	0.6350	0.7567
	Concatenate	RGB & Thermal	0.3976	0.6387	0.7567
	Concatenate + PCA	RGB & Thermal	0.3980	0.6373	0.7681
Late fusion	Intensity-based fusion	RGB & Thermal	0.3870	0.6324	0.7796
	Tracking-quality-based fusion	RGB & Thermal	0.3647	0.6480	0.7454
	MFNet (Ours)	RGB & Thermal	0.4116	0.6465	0.7835

TABLE VIII

SPEED ANALYSIS FOR EACH COMPONENT IN JMMAC

Module	JMMAC(B)	MFNet	CME	TMP	BR	JMMAC
Time (sec.)	0.1160	0.0124	0.0551	0.0737	0.0415	0.2664

the affine transformation performs best and serves as our final model in camera motion estimation module.

Speed Analysis. We conduct speed analysis for JMMAC and show the average time cost for each component in Table VIII. Our tracker approximately runs at 4 FPS and the main time cost is from the appearance tracker (JMMAC(B)) with the deep feature. Our proposed MFNet is efficient to fuse the multi-modal information and provide a final response and the target motion prediction(TMP) and camera motion estimation (CME) mechanisms do not bring significant speed decline. The box regression with a real-time YOLOv2 is applied to refine the bounding box, whose computation can be negligible.

V. CONCLUSION

In this study, we propose a novel JMMAC method for robust RGB-T tracking. Our method effectively exploits both appearance and motion cues in dealing with the RGB-T tracking task. For the appearance information, we develop a novel MFNet to infer the fusion weight maps of both RGB and thermal modalities, resulting in a much reliable response map and a robust tracking performance. The experiments demonstrate that our MFNet is not only suitable for improving the tracking accuracy but also competitive in handling the image fusion task. For the motion information, we attempt to jointly consider camera motion and target motion, enabling the tracker to become much more robust when the appearance cue is unreliable. Extensive results on GTOT, RGBT234, and VOT19-RGBT datasets show that the proposed JMMAC tracker achieves remarkably better performance than other state-of-the-art algorithms.

REFERENCES

- [1] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
- [2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 850–865.
- [3] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4787, May 2017.
- [4] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.
- [5] C. O. Conaire, N. E. O'Connor, E. Cooke, and A. F. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *Proc. 9th Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–7.
- [6] C. Ó. Conaire, N. E. O'Connor, and A. Smeaton, "Thermo-visual feature fusion for object tracking using multiple spatiogram trackers," *Mach. Vis. Appl.*, vol. 19, nos. 5–6, pp. 483–494, Oct. 2008.
- [7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
- [8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6638–6646.
- [9] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.
- [10] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.
- [11] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, Jan. 2014.
- [12] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, "Deep adaptive fusion network for high performance RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.
- [13] F. Gustafsson *et al.*, "Particle filters for positioning, navigation, and tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 425–437, Aug. 2002.
- [14] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
- [15] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
- [16] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.
- [17] M. Kristan *et al.*, "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2019, pp. 1–36.
- [18] G. Y. Kulikov and M. V. Kulikova, "The accurate continuous-discrete extended Kalman filter for radar tracking," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 948–958, Feb. 2016.
- [19] B. K. S. Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," *Signal, Image Video Process.*, vol. 7, no. 6, pp. 1125–1143, Nov. 2013.
- [20] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, Jul. 2015.
- [21] J. Kwon, H. S. Lee, F. C. Park, and K. M. Lee, "A geometric particle filter for template-based visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 625–643, Apr. 2014.
- [22] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.
- [23] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, no. 12, Dec. 2019, Art. no. 106977.
- [24] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

- [25] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1856–1864.
- [26] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 808–823.
- [27] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, "Learning local-global multi-graph descriptors for RGB-T object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2913–2926, Oct. 2019.
- [28] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
- [29] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4140–4146.
- [30] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1728–1740, Oct. 2008.
- [31] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [32] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017.
- [33] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Sci. China Inf. Sci.*, vol. 55, no. 3, pp. 590–599, Mar. 2012.
- [34] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
- [35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [36] A. Lukežić, L. Čehovin Zajc, T. Vojíř, J. Matas, and M. Kristan, "FuCoLoT—A fully-correlational long-term tracker," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 595–611.
- [37] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
- [38] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017.
- [39] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
- [40] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2015, pp. 2130–2134.
- [41] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.
- [42] W. Zheng and S. M. Bhandarkar, "A boosted adaptive particle filter for face detection and tracking," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 28–39.
- [43] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6548–6552.
- [44] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, no. 22, pp. 50–59, Jan. 2016.
- [45] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, Mar. 2002.
- [46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [47] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.
- [48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
- [49] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault diagnosis," *Knowl.-Based Syst.*, vol. 119, pp. 200–220, Mar. 2017.
- [50] I. Talmi, R. Mechrez, and L. Zelnik-Manor, "Template matching with deformable diversity similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 175–183.
- [51] O. R. Terrades, E. Valveny, and S. Tabbone, "Optimal classifier fusion in a non-Bayesian probabilistic framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1630–1644, Sep. 2009.
- [52] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.
- [53] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.
- [54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [55] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, Tech. Rep. TR 95-041, 1995.
- [56] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive Kalman filter," *J. Vis. Commun. Image Represent.*, vol. 17, no. 6, pp. 1190–1208, Dec. 2006.
- [57] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
- [58] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
- [59] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
- [60] L. Yang, B. L. Guo, and W. Ni, "Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform," *Neurocomputing*, vol. 72, nos. 1–3, pp. 203–211, Dec. 2008.
- [61] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang, "Learning target-oriented dual attention for robust RGB-T tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1–8.
- [62] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
- [63] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGB-T tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.
- [64] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Opt. Eng.*, vol. 52, no. 5, May 2013, Art. no. 057006.
- [65] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
- [66] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.
- [67] Y. Zhu, C. Li, Y. Lu, L. Lin, B. Luo, and J. Tang, "FAANet: Quality-aware feature aggregation network for RGB-T tracking," *CoRR*, vol. abs/1811.09855, pp. 1–11, Nov. 2018.
- [68] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for RGB-T tracking," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 465–472.

Pengyu Zhang received the B.E. degree in information engineering from Chang'an University, China, in 2016. He is currently pursuing the Ph.D. degree with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology (DUT), Dalian, China. His research interests include visual object tracking and multiple modality fusion.



Jie Zhao received the B.E. degree from Northwest A&F University, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Dalian University of Technology, China. Her research interests include object tracking, object detection, and self-supervised learning.





Chunjuan Bo received the Ph.D. degree in signal and information processing from the Dalian University of Technology, China, in 2019. She is currently an Associate Professor with the College of Information and Communication Engineering, Dalian Minzu University. Her research interests include image classification and object tracking.



Huchuan Lu (Senior Member, IEEE) received the M.S. degree in signal and information processing and the Ph.D. degree in system engineering from the Dalian University of Technology (DUT), China, in 1998 and 2008, respectively. He has been a Faculty (since 1998) and a Professor (since 2012) with the School of Information and Communication Engineering, DUT. His research interests include the areas of computer vision and pattern recognition. In recent years, he focuses on visual tracking and segmentation. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Dong Wang (Member, IEEE) received the B.E. degree in electronic information engineering and the Ph.D. degree in signal and information processing from the Dalian University of Technology (DUT), Dalian, China, in 2008 and 2013, respectively. He is currently a Professor with the School of Information and Communication Engineering, DUT. His research interest includes object detection and tracking.



Xiaoyun Yang received the Ph.D. degree in computing from the University of Ulster in 2006. He is currently the Director of Remark AI UK Limited. His research interests include applying computer vision research into real-world applications, recently focusing on the area of object detection/tracking/recognition and semantic segmentation.