MIRNET: A ROBUST RGBT TRACKING JOINTLY WITH MULTI-MODAL INTERACTION AND REFINEMENT

Ruichao Hou, Tongwei Ren(✉), *Gangshan Wu*

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
rc_hou@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

ABSTRACT

RGBT tracking attempts to design a robust all-weather tracker by integrating the complementary features of visible and thermal spectrums. To explore the latent interdependencies across modalities, we propose a novel real-time tracker named MIRNet, which contains a multi-modal interaction module (MIM) and a refinement mechanism (RM), thereby adaptively merging multi-modal features and achieving precise scale estimation. Specifically, to enhance instance representation in low-quality modality, the MIM reinforces discriminative features from one modality to another in a bidirectional way. Considering the negative effects of unreliable modality, we further introduce a gate function in MIM to filter redundancy. To address the problem of random drifting and estimate the precise scale in the online tracking, we present a well-designed RM that combines optical flow and refinement network. Comprehensive experiments on two public RGBT benchmarks validate that our tracker outperforms the state-of-the-art methods.

Index Terms— RGBT tracking, multi-modal interaction, gate function, optical flow, refinement network

1. INTRODUCTION

RGBT tracking aims to fuse dual-modal complementary cues while suppressing interference, and predict the bounding box of the object in the subsequent frames according to the label of the initial frame [1]. One common type of RGBT tracking is committed to yielding a comprehensive representation by integrating the advantages of heterogenous modalities. In previous works [2, 3], cross-modal ranking approach play a critical role in feature fusion. However, those methods are designed on the basis of handcraft features, which limits tracking performance. Several latest trackers seek to construct instance representations by merging the hierarchical features [4, 5, 6, 7], and introducing additional subnetworks [8, 9], to name a few. Although the aforementioned methods have effectively improved the tracking performance, their model structures are more complex. Another kind of RGBT tracking approaches focuses on generating fusion weights to measure the reliability of modality. As represented by those that leverage attention mechanisms [10, 11] or discriminative learning framework [12] , they first evaluate the modal contributions and then adopt contribution scores to guide features fusion. As shown in Fig. 1, the above trackers fail to exploit the latent inter-modal information and their performance will degrade in challenging scenarios with illumination variations and dynamic disturbance. In addition, the common challenges in visual tracking also require more attention. We observe that camera motion and scale changes lead to drifting and in this case numerous trackers are not capable of predicting the precise bounding box. Hence, those studies still leave some room for improvement in the aspects of multi-modal interaction and bounding box refinement.

Fig. 1. Visual example of comparison between our MIRNet and other trackers. Compared with these methods, our tracker achieves satisfactory results in various scenarios, such as partial occlusion, camera motion and extreme illumination.

To handle the aforementioned issues, we propose a novel tracker named MIRNet, which focuses on mining the inter-modal correlation to modulate deep features. CMPP [5] is a powerful tracker, which generates intra-modal correlations based on self-attention and merges them to enhance dual modalities. Unlike it, we design a multi-modal interaction module (MIM) to enhance target appearance in a bidirectional way, motivated by the cross-attention of the transformer [13] and the channel-wise attention [14]. The cross-modal attention views the inter-modal correlation as the attention matrix to guide feature transferring between modalities, which fully exploits the complementarity and improves the quality of both patterns. To further control multi-modal information flow, we

(✉)Corresponding Author

embed a simple and efficient gate function into MIM.

In the online tracking phase, when the object is beyond the searching region, Li *et al*. [15] endeavor to expand the search range, and Zhu *et al*. [16] propose a global re-search approach. However, these methods introduce additional interference, which exacerbates computational complexity. To this end, we develop a refinement mechanism (RM) by incorporating a fast optical flow algorithm [17] and a box refinement network. In particular, the RM is a multi-stage optimization strategy, which determines whether to switch optical flow or refinement network according to the confidence score and motion offset. Note that we select the most reliable modality as input depending on the weight score from the gate function.

The major contributions of our work are summarized as follows. (1) We propose a MIM component to reinforce instance representation via cross-modal attention mechanism and an efficient gate function. (2) We present an elaborate RM component that incorporates fast optical flow and box refinement network and realizes misalignment prevention to boost tracking performance. (3) Experiments verify that our method achieves satisfactory performance compared against the state-of-the-art trackers on two RGBT benchmarks.

2. RELATED WORK

2.1. RGBT tracking

RGBT trackers mainly include two basic network structures, one type of which is inspired by the idea of tracking by detection [18]. For instance, Wang *et al*. [5] presented a CMPP tracker to enhance feature, which aggregated instance patterns across modalities on the spatial-temporal domain. Xu *et al*. [6] designed a cross-layer bilinear pooling network with a novel multi-scale attention mechanism to achieve hierarchical feature fusion, and yielded satisfactory results. Zhu *et al*. [7] proposed a trident fusion framework to merge all the hierarchical multi-resolution features and prune redundant channels, which could avoid network overfitting and guarantee generalization ability. To boost the tracking performance in complex scenarios, some modified versions [9, 8] attempt to develop challenge attribute-aware sub-networks. Another type of trackers adopt the Siamese network as a core structure to achieve real-time performance. Zhang *et al*. [19] first trained an end-to-end RGBT tracker on the basis of DiMP and a synthetic dataset. Guo *et al*. [20] performed a response-level fusion tracking on the siamese network, whose major contribution is the deployment of weight distribution via a joint channel attention module. Commonly, the siamese-based tracker requires extra data or synthetic data for training.

2.2. Multi-stage tracking strategy

The multi-stage tracking strategy first coarsely predicts the target location and then refines the bounding box, which is a critical step to promote tracking accuracy. Yan *et al*. [21] proposed a scale estimation network consisting of a pixel-wise correlation layer and a spatial aware non-local layer that can be flexibly applied to various trackers. mfDiMP [19] employed an IoU-Net to narrow the gap between tracking results and ground truth during online tracking. Zhang *et al*. [9] leveraged target appearance and motion cues to construct a tracker for the estimation of the potential region. Although multi-stage refinement methods boost tracking performance, some specific problems in RGBT tracking are neglected, such as weak registration and high computational costs.

3. METHODOLOGY

3.1. Network architecture

The whole framework of our method is shown in Fig. 2. Our MIRNet is designed on the basis of the RT-MDNet [18]. Following the baseline, we choose a lightweight VGG-M network with atrous convolution as the backbone and expand it to a shared-parameters two-stream structure. We utilize the first three convolution layers to extract features and then use MIM that includes cross-modal attention and a gate function to enhance instance representation. The positive and negative samples are croped by the RoIAlign. To predict the score of the candidate region and obtain a coarse location, the samples are then fed into the binary classifier. Finally, we apply the RM to refining tracking results via optimizing the bounding box or estimating motion offset.

3.2. Multi-modal interaction module

Cross-modal attention. Inspired by the transformer [13], we construct a multi-head cross-modal attention to explore potential inter-modal relevance and sense global information, thereby guiding one modality to receive discriminative features from another modality. Note that it is a bidirectional learning process to improve the quality of dual-modal deep features and enhance the instance representation. The flow chart of the cross-modal attention is shown in Fig. 3.

Following the definition of self-attention [13], we use *Query*, *Key*, *Value* to construct the cross-modal attentions, all of which are obtained through 1×1 convolution and reshape operation:$(\mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{HW\times C})$. The discriminative features transformation from Thermal pattern to RGB pattern is denoted as F_{T-R}. The formulas are as follows:

$$\begin{aligned} F_{T-R}(X^R, X^T) = &\ \text{Attention}\left(Q^R, K^T, V^T\right) \\ = &\ \text{softmax}\left(\frac{Q^R\left(K^T\right)^{\mathrm{T}}}{\sqrt{d_k}}\right)V^T, \end{aligned} \tag{1}$$

where $\{X^R, X^T \in \mathbb{R}^{H\times W\times C}\}$ represents a pair of deep feature maps, and d_k is scaled factor. We establish a cross-modal correlation between Q^R and K^T, and then produce the attention matrix via Softmax$(\cdot)$. The cross-modal features are generated by weighting V^T. To learn various attention distributions, we extend cross-modal attention to a multi-head structure.

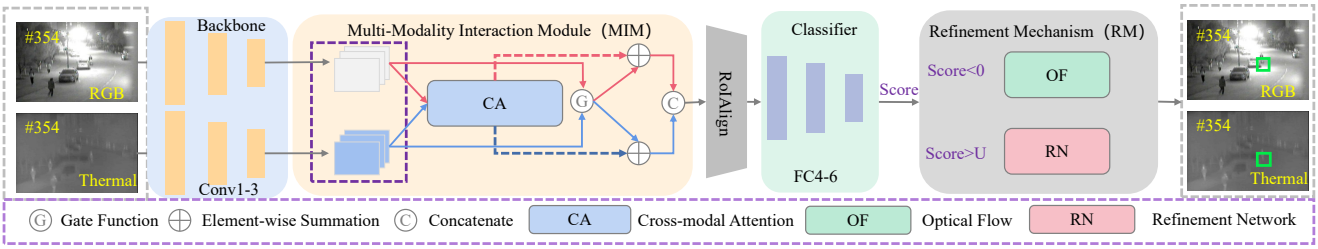

Fig. 2. Pipeline of our MIRNet, which contains two key components, namely a multi-modal interaction module and a refinement mechanism. Note that Score denotes confidence score, which is used to choose optical flow or refinement network.

$$\text{MultiHead}\left(Q^R, K^T, V^T\right) = \text{Concat}\left(H_1, ..., H_n\right) W^O, \tag{2}$$

$$H_i = \text{Attention}\left(Q^R W_i^Q, K^T W_i^K, V^T W_i^V\right), \tag{3}$$

where $W_i^Q, W_i^T, W_i^V \in \mathbb{R}^{C\times C/n}$, and $W^O \in \mathbb{R}^{C\times C}$ are weights and n means the number of head. In constrast, F_{R-T} is discriminative feature generated from RGB pattern to Thermal pattern, which can be easily calculated in the same way.

Gate function. Previous methods attempt to directly concatenate or add the multi-modal features to infer the global channel attention score but may ignore the intra-modal correlation. To this end, we propose a novel gate function to control the two patterns of information flow and establish a long-range channel dependency via cross-modal channel-wise attention. We first separately calculate the channel attention of the two modalities and then concatenate the attention vectors. Finally, the softmax$(\cdot)$ is utilized to reweight the attention scores. The structure of the gate function is shown in Fig. 4. The gate function is defined as follows.

$$S^i = \sigma\left(f_{conv}^i\left(GAP\left(X^i\right)\right), i = R, T, \right. \tag{4}$$

where $GAP(\cdot)$ denotes the global average pooling, f_{conv} represents a 1×1 convolution layer, σ is Sigmoid$(\cdot)$ function, S means the channel attention scores.

$$W^R = \text{softmax}\left(\text{Concat}\left(S^R, S^T\right)\right), \tag{5}$$

$$\bar{X}^R = X^R \cdot W^R, \bar{X}^T = X^T \cdot \left(1 - W^R\right), \tag{6}$$

where W^R is the reconstructed weight of RGB pattern, $\{\bar{X}^R, \bar{X}^T\}$ is a pair of enhanced features.

$$\widetilde{X}^R = \bar{X}^R + F_{T-R}, \widetilde{X}^T = \bar{X}^T + F_{R-T}. \tag{7}$$

We adopt residual connection to produce robust features $\widetilde{X}^R$ and $\widetilde{X}^T$, and then concatenate them to generate the high-quality fused instance representation.

3.3. Refinement mechanism

When faced with a sudden drifting between adjacent frames, the local search strategy of the base tracker is ineffective. In addition, because base tracker learns bounding box regression on the initial frame, it is hard to adaptively adjust the scale of the target in subsequent frames.

To address this challenge, we propose a refinement mechanism that involves multiple stages of optimization. In the stage of coarse location, we introduce a relocation rule on the

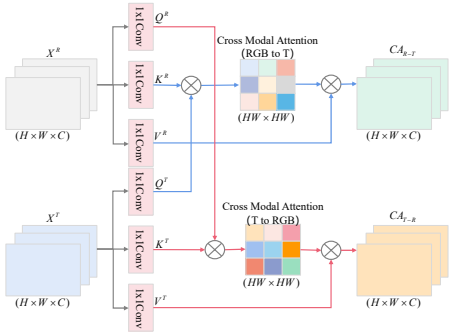

Fig. 3. Flow chart of the cross-modal attention. Herein, $\otimes$ denotes matrix multiplication.

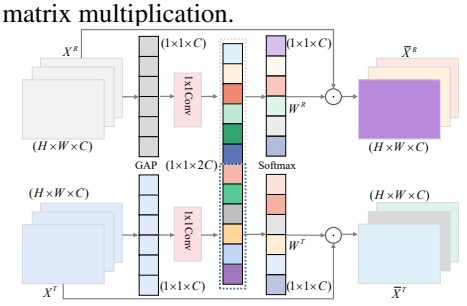

Fig. 4. Structure of the gate function. Herein, $\odot$ denotes element-wise product.

basis of the fast optical flow [17] to estimate the motion cues of key points between adjacent frames and eliminate the displacement error. Considering the stability of the key points, we choose the thermal pattern that is insensitive to illumination as input. When the confidence score is lower than 0, the target may be lost, and in this case the relocation rule will be used to calculate the motion offset $[dx, dy]$. If the offset is less than the threshold T, we reckon the local search with Gaussian sampling is able to capture the target, otherwise the candidate region will be optimized via adding offset.

In the stage of precise positioning, we embed the box refinement network into the tracking framework. The Alpha Refine [21] that consists of a pixel-wise correlation layer and a spatial aware non-local layer is viewed as a plug-and-play component. However, the pre-trained model is trained on RGB pattern, which is unsuitable for RGBT tracking and needs to be fine-tuned. We observe that it is unnecessary to consistently optimize the box because it not only increases

the time cost but also weakens the robustness when the tracking is unstable. Hence, we first train refinement network on the RGBT datasets and then refine candidates with confidence scores greater than U. Specifically, according to the weights calculated by the gated function, the most reliable pattern is fed into the refinement network. The refinement network effectively mitigates the effects of weak registration, since it is optimized on a single modality. We set T to 30 and U to 10. The analysis of parameters setting is given in Section 5.3.

4. IMPLEMENTATION DETAILS

In the offline training phase, we choose the VGG-M as the backbone and follow a multi-domain learning strategy. We then randomly select 8 frames and extract 32 positive samples and 96 negative samples by Intersection over Union (IoU) and Gaussian distribution from each frame to form a mini-batch. The AdamW [22] algorithm is applied to optimizing our network. The learning rate of the convolutional layers and the fully connected layers are set to $1e^{-4}$ and $1e^{-3}$ respectively. The heads of cross-modal attention are set to 2, and the epoch is set to 200. Note that the mixed loss function [18] consists of binary classification loss and instance embedding loss.

In the online training phase, the last FC layer needs to be reinitialized. We crop 500 positive samples and 5000 negative samples according to the label of the first frame. We then set the learning rate of the last FC to $1e^{-4}$ and the rest to $1e^{-3}$, and fine-tune all the FC layers by 50 epochs. In the tracking phase, we maintain a sample set with 256 candidate regions on the $t-1$-th frame to predict the result of t-th frame. Especially, we select the top-5 candidates with the highest confidence score, and obtain the tracking result by their average. Similarly, the long and short-term update mechanisms are used to update tracker. When the confidence score is lower than 0, the optical flow will calculate the offset, and if motion offset exceeds the threshold T, the position will need to be corrected. When the confidence score is higher than U, the refinement network is utilized to optimize the bounding box.

5. EXPERIMENTS

5.1. Datasets and metrics

In our work, comparative experiments are conducted on two public RGBT datasets, namely GTOT [24] and RGBT234 [1]. Specifically, GTOT contains 50 RGBT sequences, and RGBT234 consists of 234 RGBT sequences with 12 challenging attributes. To test the metrics on the RGBT234, our network is trained on GTOT. When we evaluate the GTOT, our tracker is trained on the entire RGBT234. For a fair comparison, we employ two classical metrics, Precision Rate (PR) and Success Rate (SR), to measure the effectiveness of trackers. Due to the difference in image resolution, we set the threshold of GTOT to 5 pixels and the threshold of RGBT234 to 20 pixels.

5.2. Comparison with the state-of-the-art

We compare our tracker with 8 high-performance methods, i.e., MANet++ [4], DAPNet [23], NRCMR [3], CAT [8], MaCNet [10], ADRNet [9], CBPNet [6], TFNet [7].

Overall performance. As reported in Table 1, our tracker outperforms other competitors with 81.6% and 58.9% in PR and SR on RGBT234. Compared to the latest ADRNet, our MIRNet achieves 0.7%, 1.8% promotion in PR and SR. Moreover, our method also achieves the best metrics with 90.7% and 74.4% in PR and SR on GTOT. These results demonstrate the effectiveness and robustness of our tracker.

Attribute-based performance. For more effective evaluation of the proposed tracker, we conduct comparative experiments under different challenging scenarios. RGBT234 covers 12 attributions, *i.e.* 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 attribute-based tracking results are shown in Table 1. We find that our tracker achieves the best performance in most scenarios. In the challenge of SV and LI, our tracker surpasses other methods, thanks to powerful instance representation and precise scale estimation. In particular, our tracker obtains gains on CM and MB attributes, which proves that the drifting issue is handled to a certain degree.

5.3. Ablation study

Single/dual-modal analysis. To demonstrate the effectiveness of multi-modal tracking, we first design a baseline named RT+RGBT, followed by RT-MDNet [18] and directly concatenate RGBT features at the third convolution layer. We then construct two variants named RT+RGB and RT+T, which denote the experiments of RT-MDNet with RGB and thermal pattern respectively. The metrics are reported in Table 2. Experiments indicate that dual-modal input effectively improves tracking performance and is superior to single-modal one.

Components analysis. To validate the effectiveness of our tracker, we construct two variants, *i.e.* 1) w/o-MIM, which prunes multi-modal interaction module , 2) w/o-RM, which prunes refinement mechanism. As can be seen from Fig. 5, the w/o-MIM framework precisely estimates the bounding box and outperforms the baseline by 1.9% and 5.1% in PR and SR. Compared to ADRNet [9], the w/o-MIM variant performs at the same level on SR. The w/o-RM variant performs as well as the hierarchical feature learning-based trackers, namely, CBPNet [6] and MaCNet [10], and has 3.1% and 2.6% promotion against the baseline in PR and SR, which indicates the MIM is an efficient component with low computational costs. The experimental results suggest that MIRNet combines the advantages of the two variants and thereby boosts the tracking performance.

Table 1. Comparison results of our method against with the state-of-the-art trackers. Attribute-based and overall performance are evaluated by PR/SR scores(%), and are produced on RGBT234 and GTOT, respectively. PR and SR denote precision rate and success rate. The best and second best results are in red and blue, respectively.

	MANet++ [4]	DAPNet [23]	MaCNet [10]	CAT [8]	NRCMR [3]	ADRNet [9]	CBPNet [6]	TFNet [7]	**MIRNet**
NO	89.8 / 65.4	90.0 / 64.4	92.7 / 66.5	93.2 / 66.8	89.0 / 60.4	91.7 / 65.8	92.0 / 64.7	93.1 / 67.3	95.4 / 72.4
PO	85.2 / 59.3	82.1 / 57.4	81.1 / 57.2	85.1 / 59.3	78.9 / 54.9	86.3 / 61.2	83.6 / 57.2	83.6 / 57.8	86.1 / 62.7
HO	70.4 / 47.1	66.0 / 45.7	70.9 / 48.8	70.0 / 48.0	59.8 / 40.9	70.8 / 49.1	69.5 / 46.4	72.1 / 49.1	71.0 / 49.0
LI	81.1 / 55.1	77.5 / 53.0	77.7 / 52.7	81.0 / 54.7	73.6 / 50.2	80.2 / 55.1	80.8 / 54.0	80.5 / 54.1	83.4 / 57.5
LR	82.3 / 54.5	75.0 / 51.0	78.3 / 52.3	82.0 / 53.9	74.9 / 47.2	83.1 / 55.6	79.8 / 52.4	83.7 / 54.4	83.9 / 56.3
TC	80.3 / 57.6	76.8 / 54.3	77.0 / 56.3	82.0 / 53.9	73.8 / 48.3	78.9 / 58.9	77.6 / 55.6	80.9 / 57.7	81.1 / 59.1
DEF	75.3 / 53.5	71.7 / 51.8	73.1 / 51.4	76.2 / 54.1	69.7 / 50.0	74.3 / 52.9	73.6 / 50.7	76.5 / 54.3	77.8 / 58.1
FM	70.0 / 45.3	67.0 / 44.3	72.8 / 47.1	73.1 / 47.0	65.9 / 40.8	77.6 / 50.3	70.8 / 44.7	78.2 / 49.0	68.3 / 47.1
SV	78.9 / 55.4	78.0 / 54.2	78.7 / 56.1	79.7 / 56.6	72.0 / 49.5	79.0 / 56.2	80.1 / 54.9	80.3 / 56.8	82.7 / 61.9
MB	72.0 / 51.1	65.3 / 46.7	71.6 / 52.5	68.3 / 49.0	62.6 / 44.2	72.7 / 53.0	70.1 / 49.5	70.2 / 50.6	74.6 / 54.6
CM	74.7 / 52.3	66.8 / 47.4	71.7 / 51.7	75.2 / 52.7	65.4 / 46.3	75.7 / 53.5	73.3 / 51.0	75.0 / 53.4	76.4 / 55.4
BC	76.7 / 49.1	71.7 / 48.4	77.8 / 50.1	81.1 / 51.9	65.5 / 41.8	78.9 / 52.7	80.6 / 50.2	81.3 / 52.5	78.9 / 51.7
ALL	80.0 / 55.4	76.6 / 53.7	79.0 / 55.4	80.4 / 56.1	72.9 / 50.2	80.9 / 57.1	79.4 / 54.1	80.6 / 56.0	81.6 / 58.9
ALL (GTOT)	90.1 / 72.3	88.2 / 70.7	88.0 / 71.4	88.9 / 71.7	83.7 / 66.4	90.4 / 73.9	88.5 / 71.6	88.6 / 72.9	90.9 / 74.4

Table 2. Single/dual-modal analysis on two benchmarks.

PR/SR(%)	RT+RGB	RT+T	RT+RGBT
RGBT234	71.6 / 49.1	65.2 / 43.4	76.8 / 52.0
GTOT	81.0 / 65.2	65.0 / 54.2	85.6 / 68.2

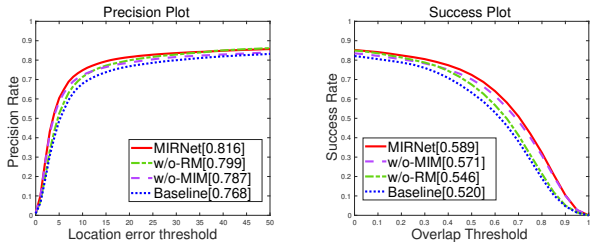

Fig. 5. Comparison of ablation experiments on RGBT234.

Parameters analysis. The offset threshold T and the confidence socre threshold U are the key parameters in online tracking. To measure the influence on performance, we set $T = \{20, 30, 40\}$ and $U = \{5, 10, 15\}$ and fix each set of parameters during the testing experiments. The results are shown in Table 3. We find that our tracker achieves the top performance with $T = 30$ and $U = 10$. When these two thresholds are too high, the RM will neglect some candidates that need to be refined, causing a decrease in robustness.

Table 3. Comparison of different parameters on RGBT234.

PR/SR(%)	$U = 5$	$U = 10$	$U = 15$
$T = 20$	81.2 / 58.7	80.6 / 57.4	81.9 / 57.1
$T = 30$	80.8 / 58.1	81.6 / 58.9	80.5 / 57.0
$T = 40$	80.4 / 57.8	81.1 / 58.4	80.5 / 56.7

5.4. Efficiency analysis

Our MIRNet is implemented on the PyTorch platform with an AMD 5600X CPU, 16GB RAM, and a NVIDIA GeForce RTX3090 GPU with 24GB memory. We analyze the running time of competitors on RGBT234, and set Frames Per Second (FPS) as evaluation metric and adopt SR to represent accuracy. Fig. 6 shows the results of efficiency analysis. Although NRCMR achieves the top speed (35FPS), the tracking accuracy is unsatisfactory. Our tracker adaptively updates the target representation, which saves the online training time and approximately reaches real-time speed (30FPS). It is makes a trade-off between efficiency and performance.

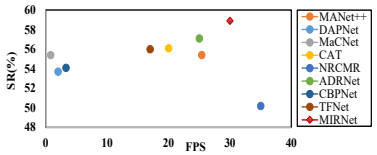

Fig. 6. Accuracy-speed plot on RGBT234.

5.5. Qualitative Analysis

We conduct qualitative comparison between MIRNet and other trackers under four challenging scenarios, and present visualization results in Fig. 7. We observe that our tracker achieves satisfactory results in complex environments, including camera motion and extreme illumination. For example, in Fig. 7(a)-(c), the target is affected by illumination, scale variations and occlusion, which pose a challenge for most competitors. However, benefiting from the robust instance representation and refinement, our tracker realizes stable tracking. Moreover, our proposed tracker has the capacity to mitigate the impact of weak registration and produce precise bounding boxes. Visualization results demonstrate the feasibility of our MIM and RM components. Furthermore, we show a failure case resulted in fast motion in Fig. 7(d).

6. CONCLUSION

In this work, we proposed a high-performance RGBT tracker called MIRNet, which contains two components, MIM and RM. To enhance instance representation and filter redundant features, a cross-modal attention and a gate function are introduced to MIM, which boosts the stability of our tracker in complex scenarios. To tackle the drifting issues, we combined the optical flow and refinement network in RM and have facilitated the regression of bounding boxes with a multi-stage

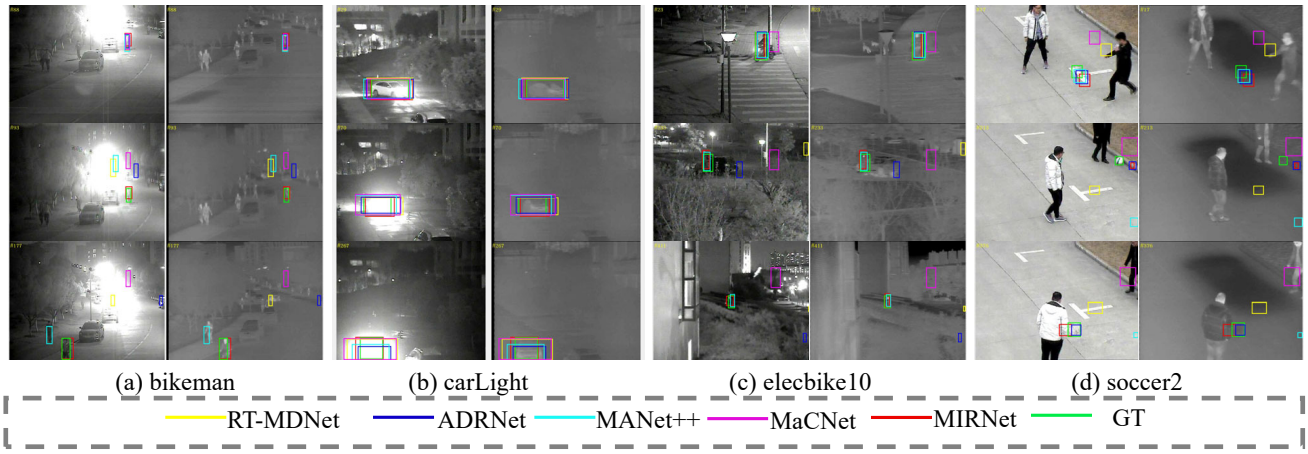

Fig. 7. Qualitative comparison between MIRNet and other trackers on four challenging sequences.

optimization strategy. Experimental results validate that our tracker achieves state-of-the-art performance on two public RGBT benchmarks while meeting real-time requirements.

Acknowledgments This work is supported by National Science Foundation of China (62072232), Natural Science Foundation of Jiangsu Province (BK20191248), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

7. REFERENCES

[1] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *PR*, 2019.

[2] 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 *ECCV*, 2018.

[3] C. Li, Z. Xiang, J. Tang, B. Luo, and F. Wang, "Rgbt tracking via noise-robust cross-modal ranking," *TNNLS*, 2021.

[4] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "Rgbt tracking via multi-adapter network with hierarchical divergence loss," *TIP*, 2021.

[5] C. Wang, C. Xu, Z. Cui, L. Zhou, T. Zhang, X. Zhang, and J. Yang, "Cross-modal pattern-propagation for rgb-t tracking," in *CVPR*, 2020.

[6] Q. Xu, Y. Mei, J. Liu, and C. Li, "Multimodal cross-layer bilinear pooling for rgbt tracking," *TMM*, 2021.

[7] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, "Rgbt tracking by trident fusion network," *TCSVT*, 2021.

[8] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware rgbt tracking," in *ECCV*, 2020.

[9] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time rgb-t tracking," *IJCV*, 2021.

[10] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in rgb-t videos using modal-aware attention network and competitive learning," *Sensors*, 2020.

[11] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust rgbt tracking," *TIV*, 2020.

[12] X. Lan, M. Ye, S. Zhang, and P. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *AAAI*, 2018.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Lukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[15] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *CVPR*, 2018.

[16] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance-specific proposals," in *CVPR*, 2016.

[17] J.Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel corporation*, 2001.

[18] I. Jung, J. Son, M. Baek, and B. Han, "Real-time mdnet," in *ECCV*, 2018.

[19] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. Shahbaz Khan, "Multi-modal fusion for end-to-end rgb-t tracking," in *ICCVW*, 2019.

[20] C. Guo, D. Yang, C. Li, and P. Song, "Dual siamese network for rgbt tracking via fusing predicted position maps," *The Visual Computer*, 2021.

[21] B. Yan, X. Zhang, D. Wang, H. Lu, and X. Yang, "Alpha-refine: Boosting tracking performance by precise bounding box estimation," in *CVPR*, 2021.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.

[23] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for rgbt tracking," in *ACM MM*, 2019.

[24] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *TIP*, 2016.