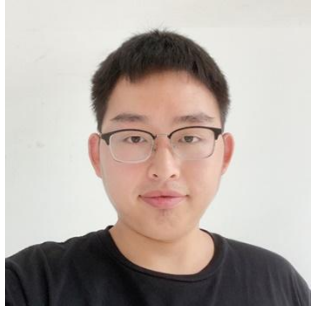
ICME2022

#1520

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

July 21, 2022

南京大學
NANJING UNIVERSITY

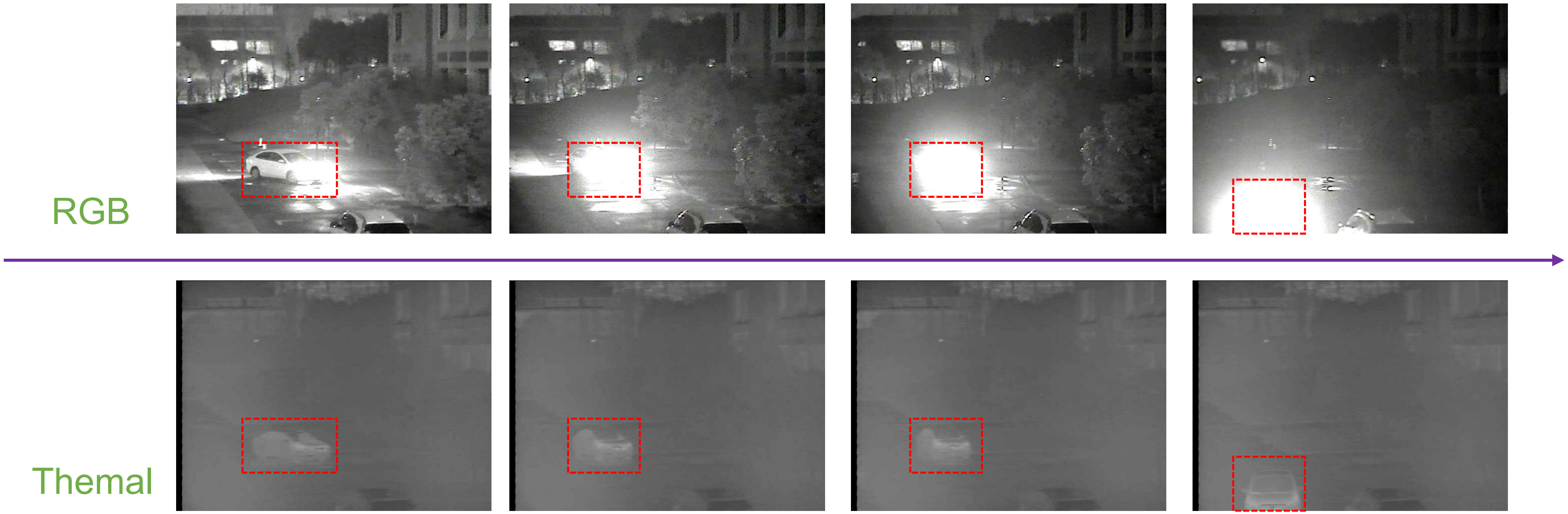
MAGUS
MediA recoGnition
and UnderStanding

Outline

- Introduction
- Methodology
- Experiments
- Conclusion

Introduction

- Task Definition: RGBT tracking is a branch of VOT, which attempts to design a robust all-weather tracker by integrating the complementary features of visible and thermal spectrums.



Introduction

- Problems

- The existing RGBT trackers are not fully exploit the latent multi-modal information.
- Camera motion, scale changes and extreme illumination lead to drifting.
- Numerous RGBT trackers are not capable of predicting the precise bounding box.

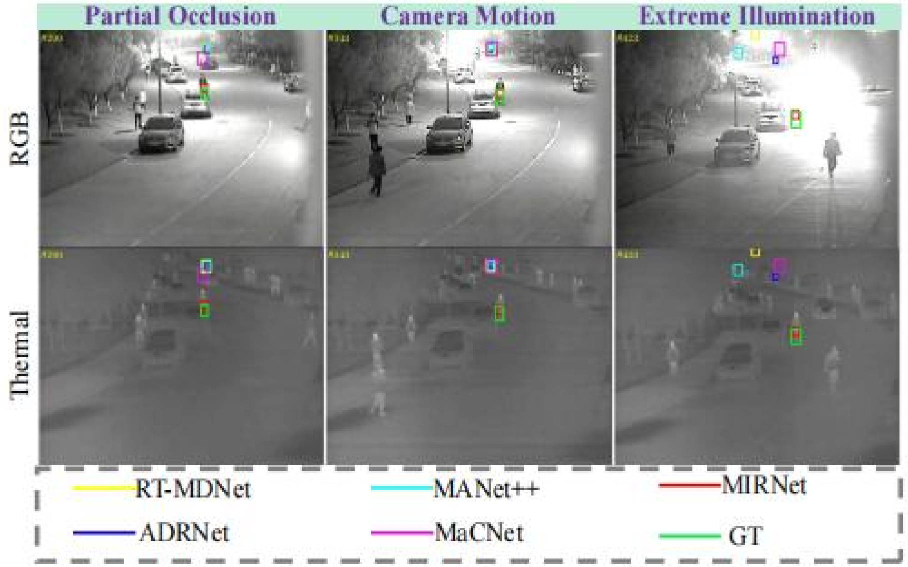


Fig. 1. Visual example of comparison between our MIRNet and other trackers. Compared with these methods, our tracker achieves satisfactory results in various scenarios, such as partial occlusion, camera motion and extreme illumination.

Introduction

- Main contributions

1. We propose a MIM component to reinforce feature representation via cross-modal attention and an efficient gate function.
2. We present an elaborate RM component that incorporates fast optical flow and box refinement network 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.

Methodology

- Network architecture

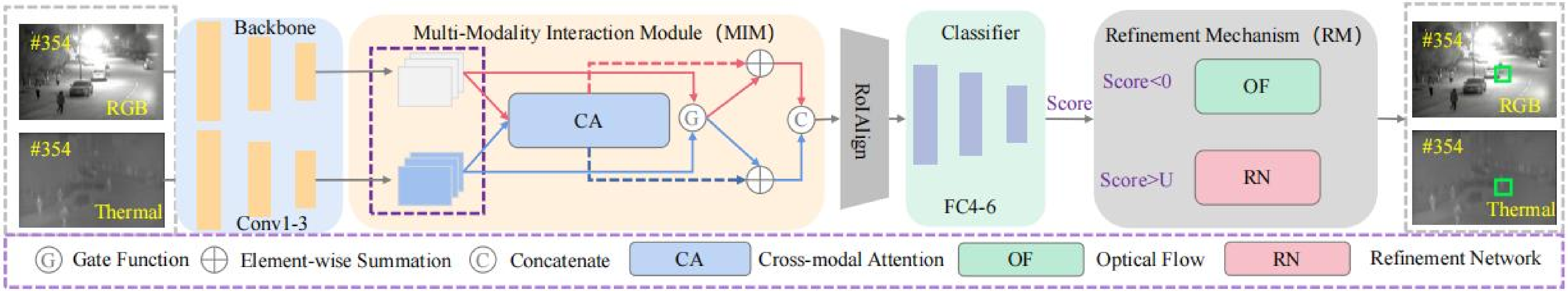
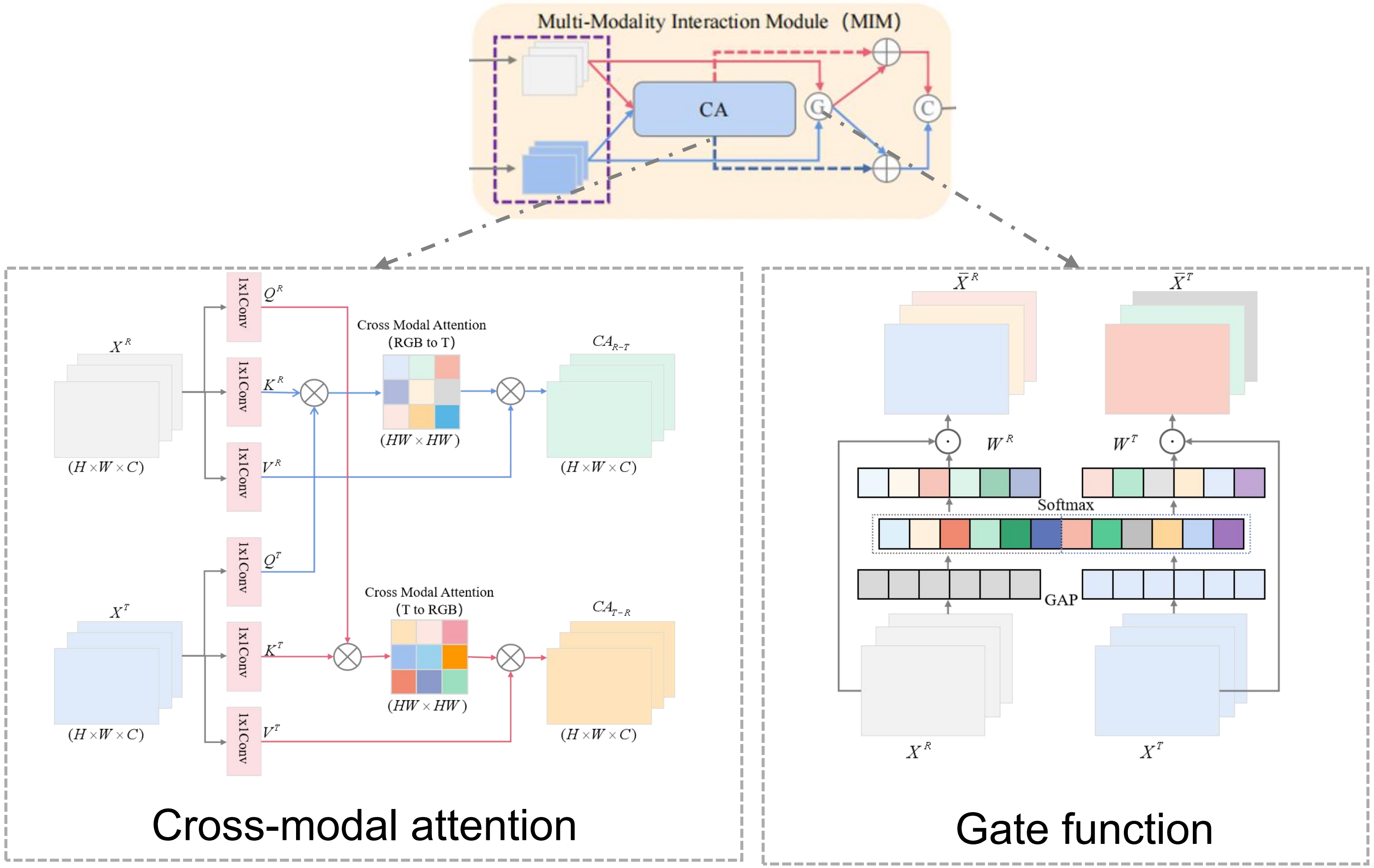


Fig. 2. Pipeline of our MIRNet, which contains two key components, namely a multi-modal interaction module and a refinement mechanism. Note that Score denotes confidence score, which is used to choose optical flow or refinement network.

Methodology

- Multi-modal interaction Module（MIM）



Cross-modal attention

Gate function

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

$$\text{MultiHead}(Q^R, K^T, V^T) = \text{Concat}\,(H_1, \ldots, H_n) W^O \quad （2）$$

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

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

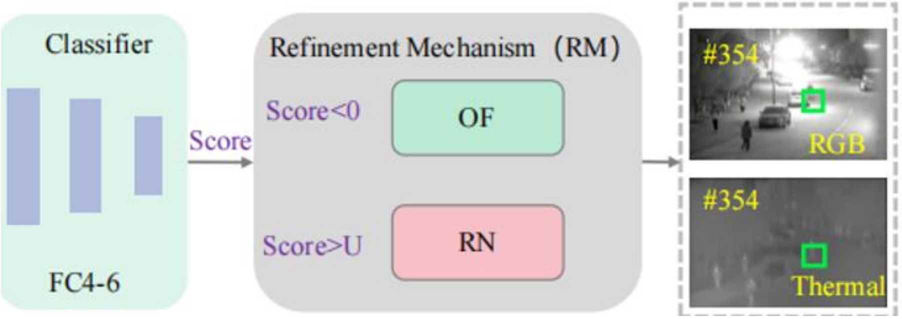
$$W^R = \text{softmax}(Concat\,(S^R, S^R)) \quad （5）$$

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

$$\tilde{X}^R = \bar{X}^R + F_{T-R}, \tilde{X}^T = \bar{X}^T + F_{R-T} \quad （7）$$

Methodology

- Refinement Mechanism （RM）



- In the stage of coarse location, when the confidence score is lower than 0, the target may be lost, and in this case the fast optical flow 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 (Alpha Refine) into the RGBT tracking framework. We refine candidates with confidence scores greater than U. According to the weights calculated by the gated function, the most reliable pattern is fed into the network.

Experiments

- Datasets
  - GTOT50
  - RGBT234

- Evaluation Metrics
  - Precision Rate (PR)
  - Success Rate (SR)

Experiments

- Comparison with the state-of-the-art

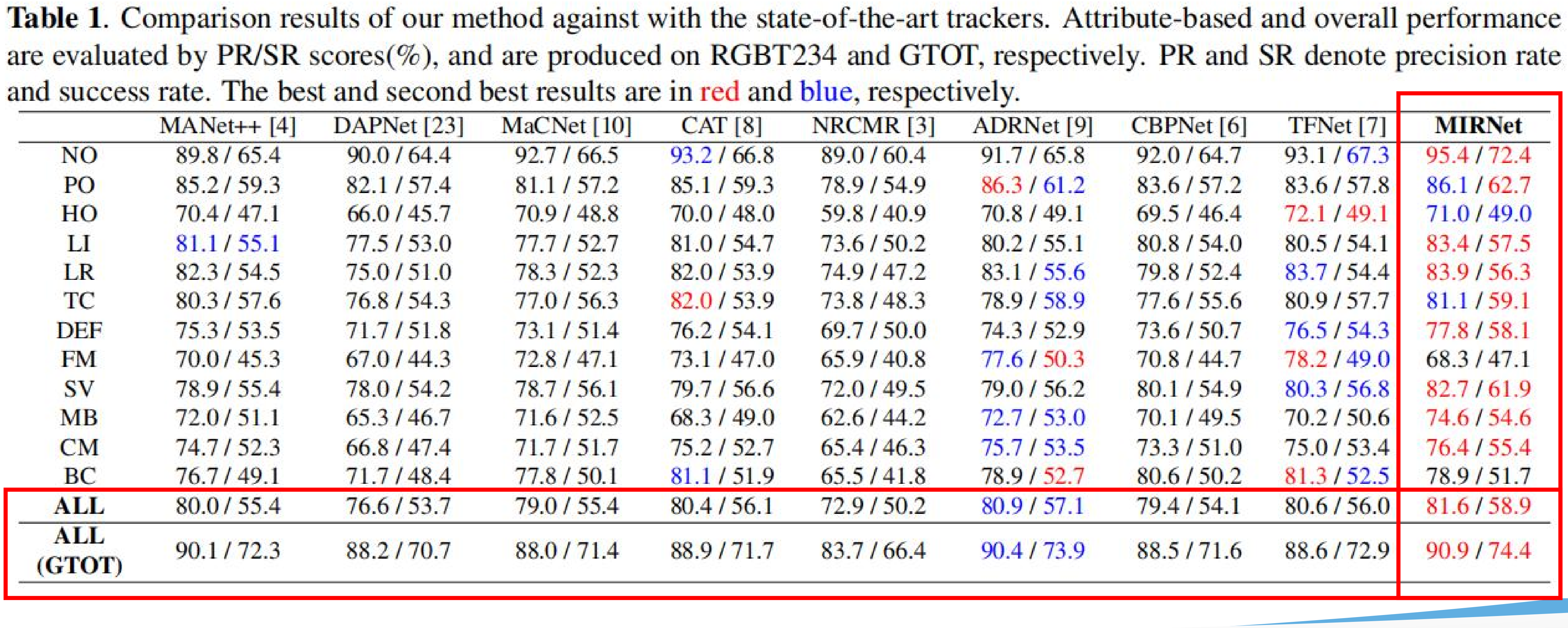


Table 1. Comparison results of our method against with the state-of-the-art trackers. Attribute-based and overall performance are evaluated by PR/SR scores(%), and are produced on RGBT234 and GTOT, respectively. PR and SR denote precision rate and success rate. The best and second best results are in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

Experiments

- Ablation study

I.Single/dual-modal analysis.

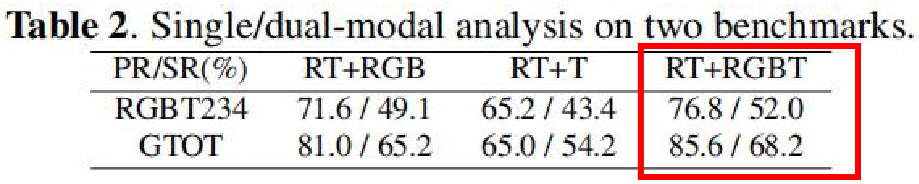


Table 2. Single/dual-modal analysis on two benchmarks.

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

II.Components analysis.

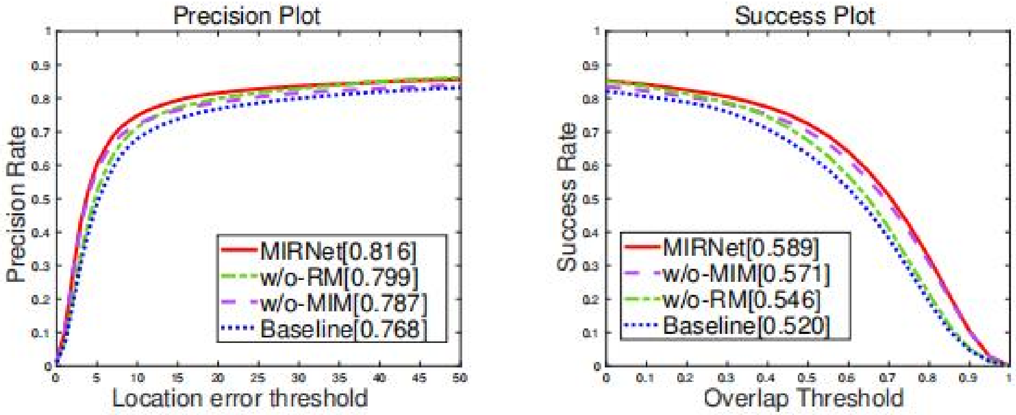


Fig. 5. Comparison of ablation experiments on RGBT234.

III.Parameters analysis.

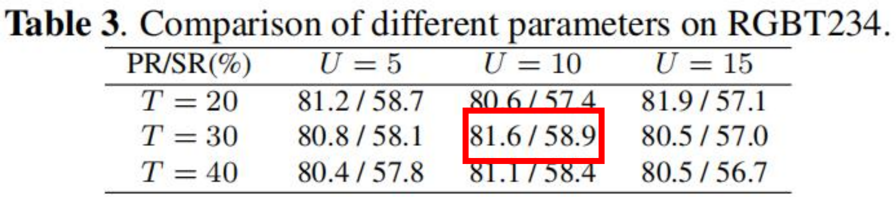
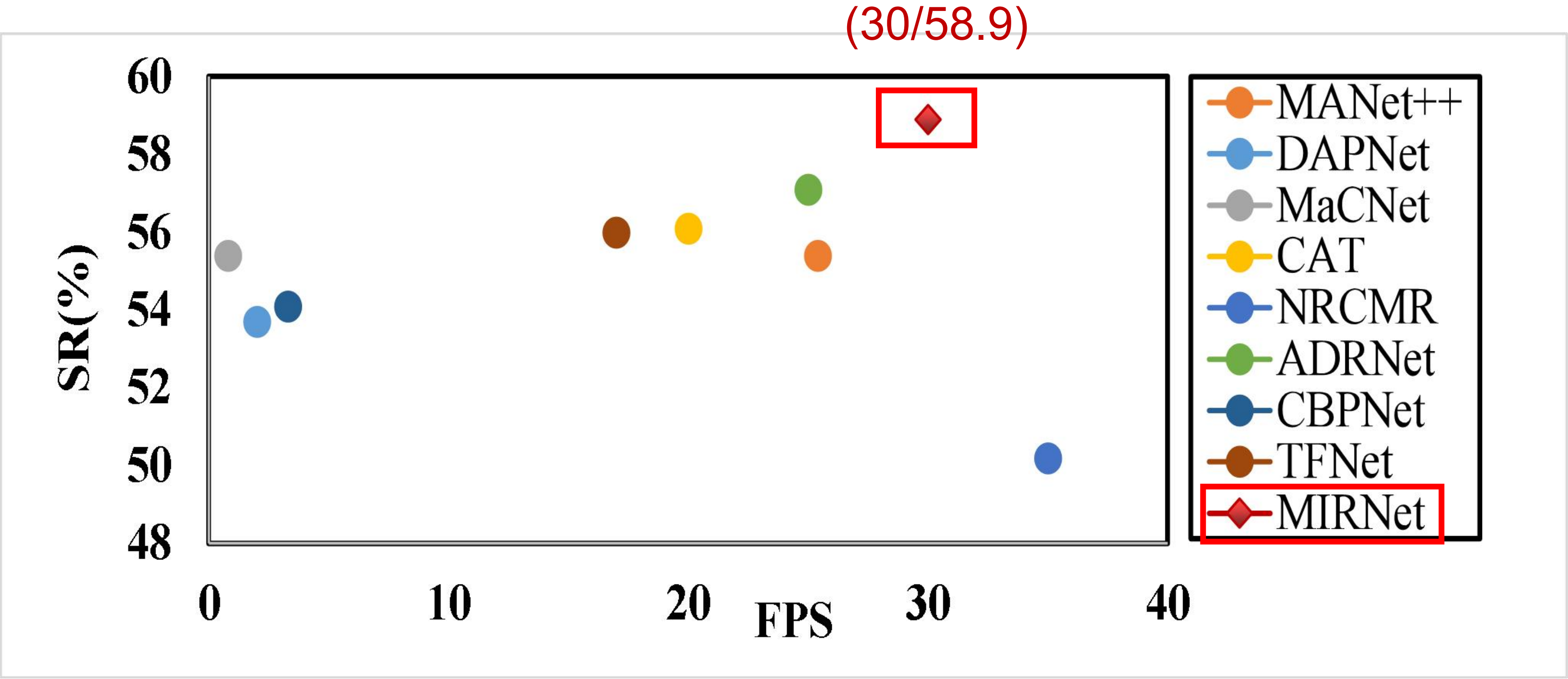


Table 3. Comparison of different parameters on RGBT234.

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

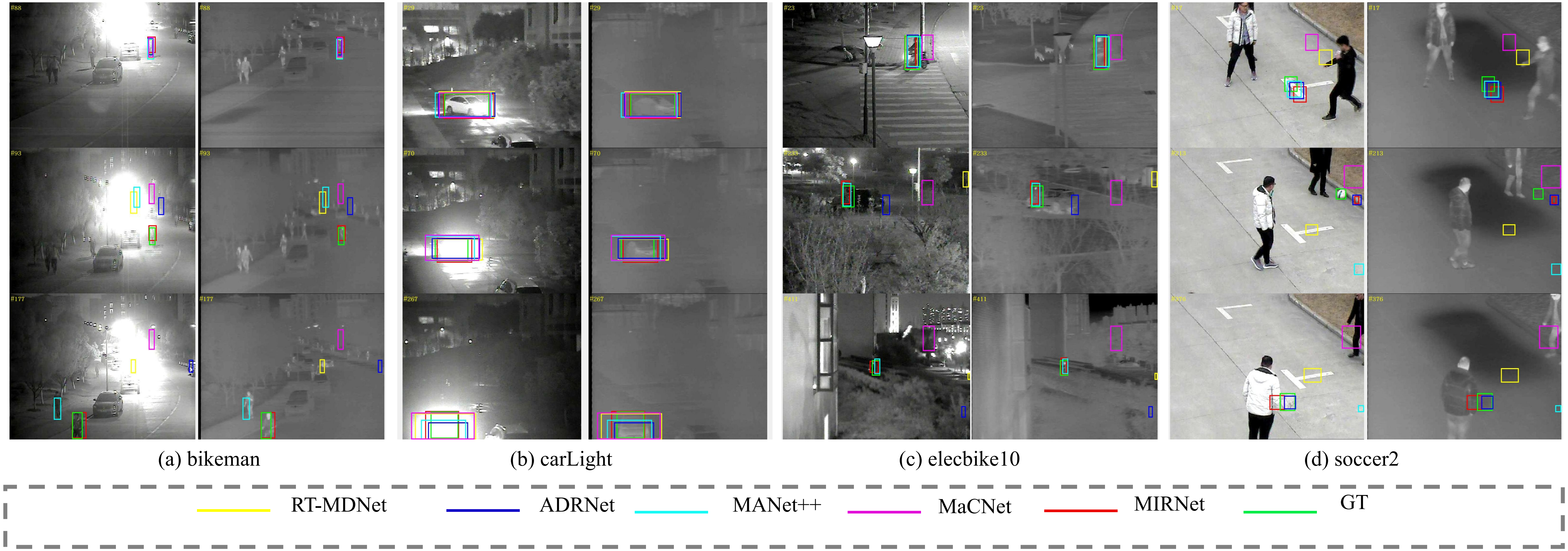
Experiments

- Efficiency analysis



Experiments

- Qualitative Analysis



(a) bikeman (b) carLight (c) elecbike10 (d) soccer2

Conclusion

- In this work, we proposed a high-performance RGBT tracker called MIRNet.
  - 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 issue, we combined the optical flow and refinement network in RM and have facilitated the regression of bounding boxes with a multi-stage optimization strategy.
  - Experimental results validate that our tracker achieves state-of-the-art performance on two public RGBT benchmarks while meeting real-time requirements.

Thank you for your attention!

rc_hou@smail.nju.edu.cn

南京大學
NANJING UNIVERSITY

MAGUS
MediA recoGnition
and UnderStanding