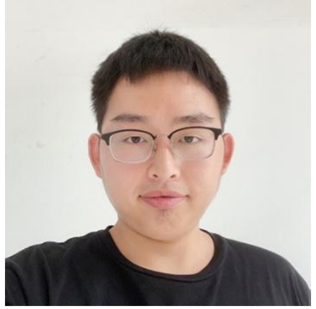
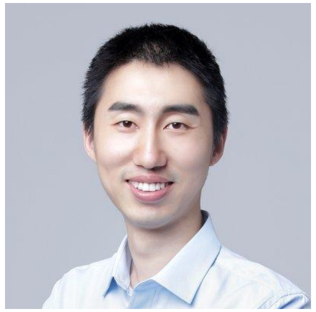
IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

MTNet: Learning Modality-aware Representation with Transformer for RGBT Tracking






Ruichao Hou Boyue Xu Tongwei Ren* Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

July 12, 2023

NANJING UNIVERSITY

MAGUS
MediA recoGnition and UnderStanding

Outline

- Introduction
- Methodology
- Experiments
- Conclusion

Introduction

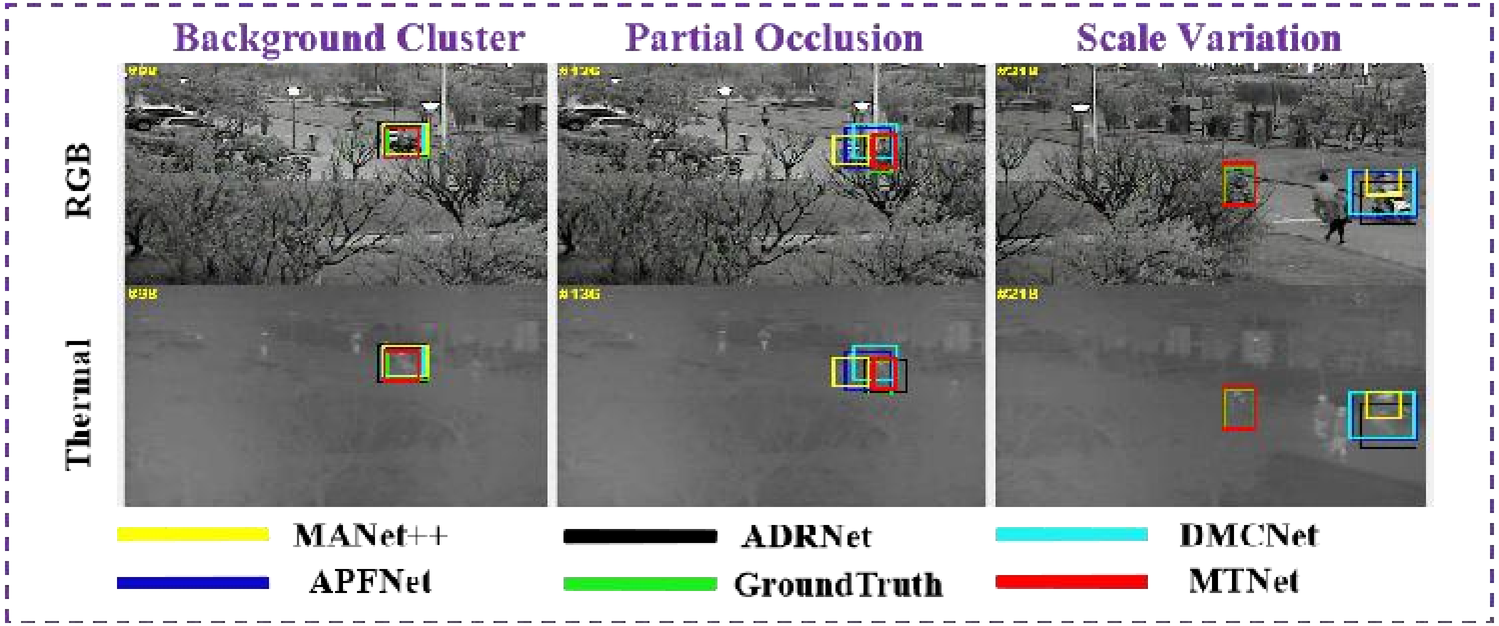
- Task Definition: RGBT tracking is a part of VOT, which attempts to design a robust all-weather tracker by integrating the complementary features of visible and thermal modality.

RGB

Themal

Motivations

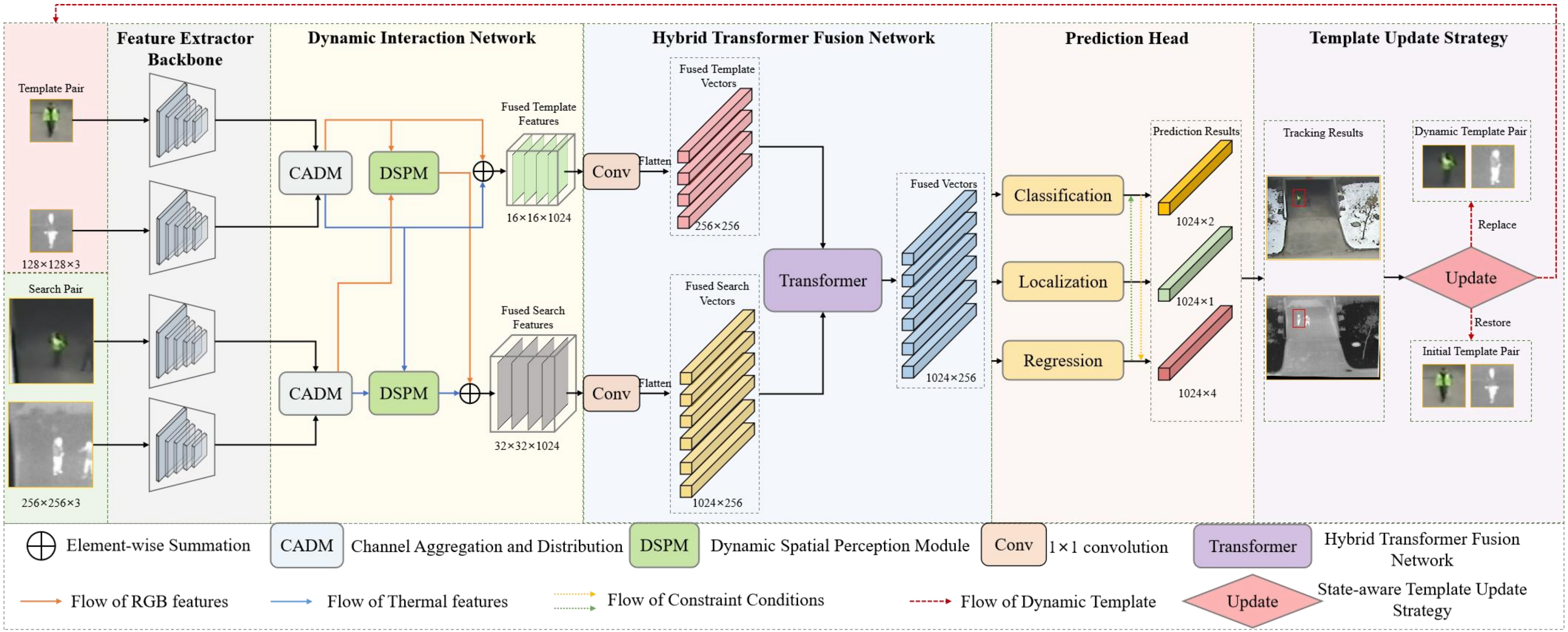
- How to efficiently extract discriminative cues from heterogeneous modalities conducive to instance representation?
- How to estimate the precise bounding box and tackle the tracking challenges?



Main Contributions

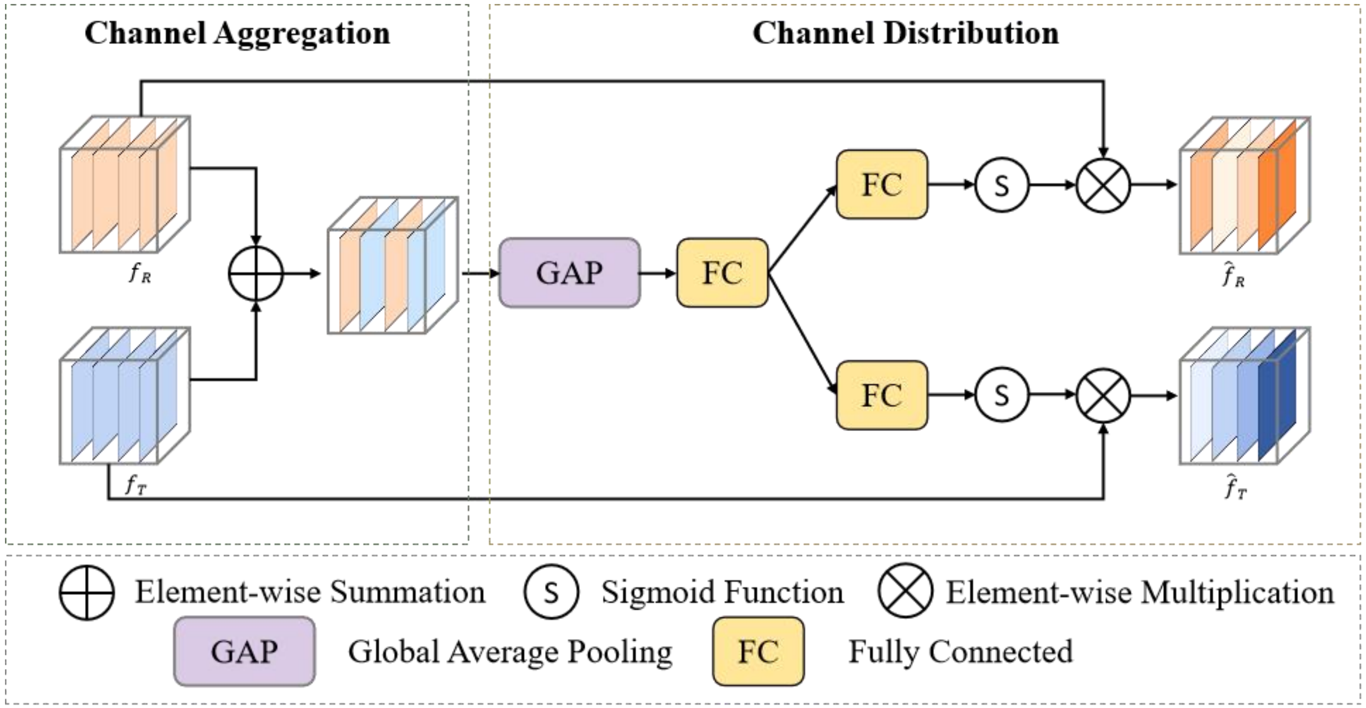
1. We propose a novel RGBT tracker that combines the locality and hierarchy of CNN and the global dependency of the transformer to learn modality-aware representations.
2. We design a trident prediction head by developing the mutual constraint loss function to improve localization accuracy. It further integrates a state-aware template update strategy to boost tracking performance.
3. Experiments verify that our method achieves satisfactory performance compared against the state-of-the-art trackers on three RGBT benchmarks.

Methodology



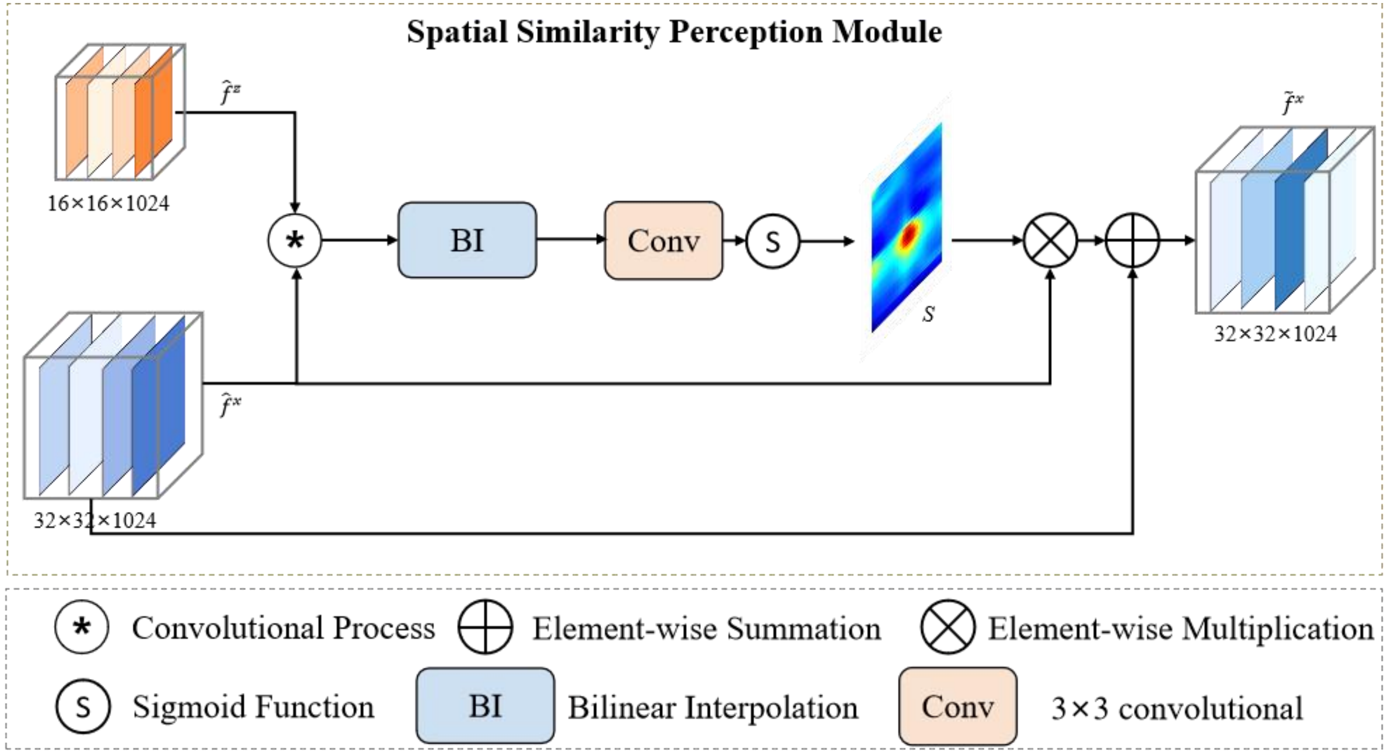
Modality-aware Network

- Channel Aggregation and Distribution Module

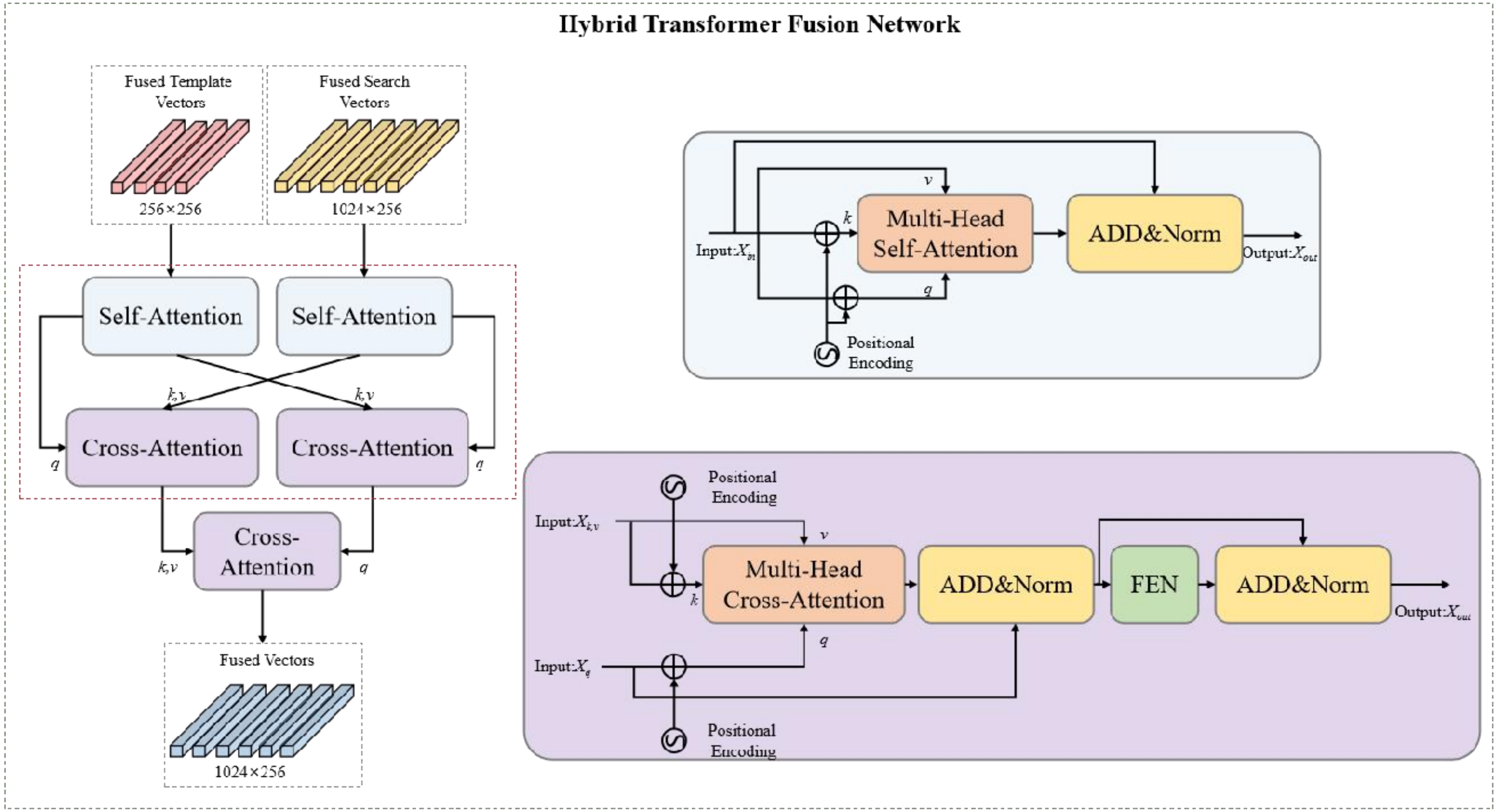


Modality-aware Network

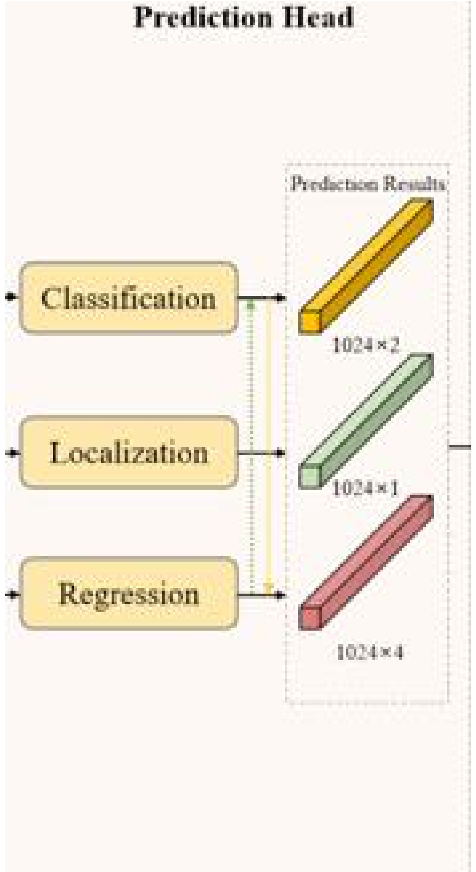
- Spatial Similarity Perception Module



Hybrid Transformer Fusion Network



Trident Prediction Heads



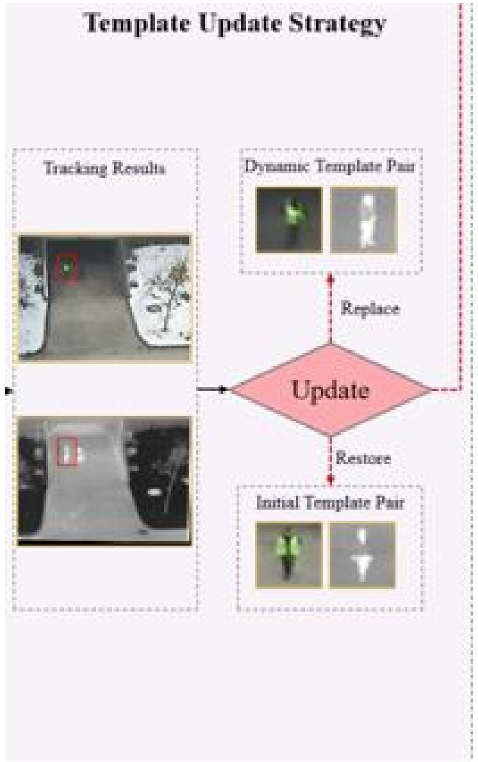
$$\mathcal{L}_{cls} = -\sum_j ((y_j \log(p_j) IoU + (1 - y_j) \log(1 - p_j))),$$

$$\mathcal{L}_{reg} = \sum_j \Pi_{y_j=1}(\lambda_1 \mathcal{L}_1(b_j, \hat{b_j}) + \lambda_C \mathcal{L}_{CIoU}(b_j, \hat{b_j}) p_j),$$

$$\mathcal{L}_{loc} = -\sum_j (O_j \log(p_j^{loc}) + (1 - O_j) \log(1 - p_j^{loc})),$$

$$\mathcal{L} = n_1 \mathcal{L}_{cls} + n_2 \mathcal{L}_{reg} + n_3 \mathcal{L}_{loc}.$$

State-aware Template Update Strategy



If the tracking template is not updated in a timely manner, tracking failures can occur. Given the real-time requirements, it is preferable to design a low-cost update strategy instead of relying on an additional auxiliary model. To achieve this, the proposed strategy divides the tracking process into three states based on confidence levels, i.e., steady state, transient steady state, and unstable state. Note that confidence is calculated by multiplying classification scores and localization scores. To pursue the best performance, we set different update intervals for each state.

Experiments

- Datasets
  - GTOT50
  - RGBT234
  - LasHeR

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

Comparison with the SOTA Trackers

TABLE I
COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART TRACKERS. ATTRIBUTE-BASED AND OVERALL PERFORMANCE ARE EVALUATED BY PR/SR SCORES(%) AND ARE PRODUCED ON RGBT234. THE BEST AND SECOND BEST RESULTS ARE IN RED AND GREEN.

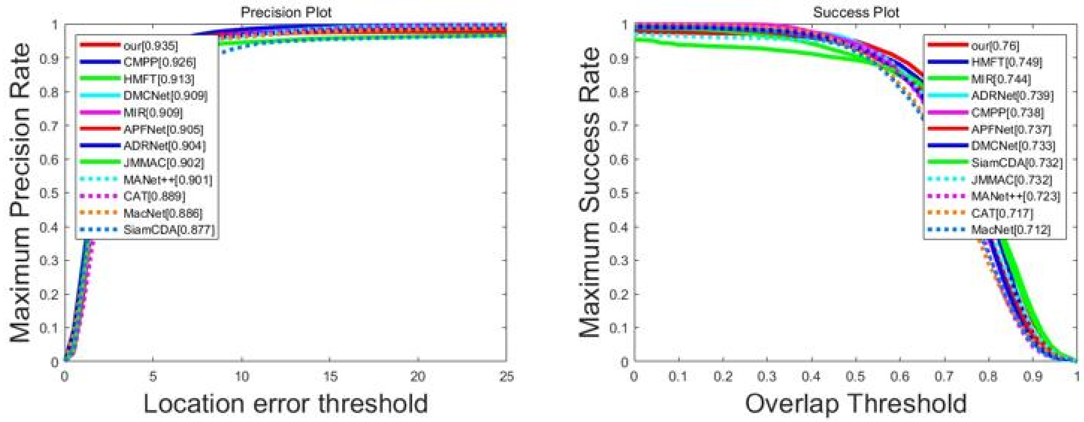
Trackers	DMCNet [5]	MIRNet [16]	APFNet [10]	AGMINet [7]	MFGNet [17]	SiamCDA [13]	RMWT [14]	HMFT [6]	**MTNet**
Pub. Info.	TNNLS2022	ICME2022	AAAI2022	TIM2022	TMM2022	TCSVT2022	KBS2022	CVPR2022	-
NO	92.3 / 67.1	95.4 / 72.4	94.8 / 68.0	94.9 / 69.1	92.0 / 64.0	88.4 / 66.4	92.1 / 70.8	90.9 / 67.4	91.0 / 67.8
PO	89.5 / 63.1	86.1 / 62.7	86.3 / 60.6	90.2 / 63.9	84.3 / 58.0	84.2 / 63.9	85.4 / 63.6	85.7 / 62.1	88.7 / 64.8
HO	74.5 / 52.1	71.0 / 49.0	73.8 / 50.7	72.9 / 50.3	66.2 / 44.3	66.2 / 48.7	75.2 / 55.5	66.4 / 46.9	78.6 / 56.3
LI	85.3 / 58.7	83.4 / 57.5	84.3 / 56.9	87.0 / 59.8	79.1 / 54.2	81.8 / 61.2	84:1 / 61.5	83.3 / 59.1	83.3 / 59.5
LR	85.4 / 57.9	83.9 / 56.3	84.4 / 56.5	86.7 / 57.2	79.3 / 49.5	70.9 / 49.9	76.6 / 55.0	76.3 / 57.1	80.4 / 55.4
TC	87.2 / 61.2	81.1 / 59.1	82.2 / 58.1	80.6 / 59.2	81.8 / 55.8	67.4 / 47.7	78.2 / 58.6	72.2 / 50.4	86.1 / 61.6
DEF	77.9 / 56.5	77.8 / 58.1	78.5 / 56.4	79.5 / 56.8	72.1 / 50.8	77.9 / 59.2	80.3 / 62.0	77.6 / 57.9	84.7 / 64.0
FM	80.0 / 52.4	68.3 / 47.1	79.1 / 51.1	79.4 / 51.2	72.5 / 44.6	61.4 / 45.3	74.3 / 55.3	65.9 / 46.9	79.2 / 58.0
SV	84.6 / 59.8	82.7 / 61.9	83.1 / 57.9	83.2 / 59.3	76.1 / 52.8	77.7 / 59.3	86.1 / 65.9	80.0 / 59.2	89.0 / 66.1
MB	77.3 / 55.9	74.6 / 54.6	74.5 / 54.5	78.2 / 57.5	73.7 / 51.0	63.6 / 47.9	76.8 / 57.8	70.6 / 50.9	83.4 / 61.6
CM	80.1 / 57.6	76.4 / 55.4	77.9 / 56.3	79.0 / 57.5	73.2 / 50.4	73.3 / 54.7	83.1 / 62.7	77.9 / 56.2	86.0 / 63.4
BC	83.8 / 55.9	78.9 / 51.7	81.3 / 54.5	83.3 / 55.3	74.3 / 45.9	74.0 / 52.9	74.5 / 52.5	73.8 / 49.8	74.9 / 50.8
ALL	83.9 / 59.3	81.6 / 58.9	82.7 / 57.9	84.0 / 59.2	78.3 / 53.5	76.0 / 56.9	82.5 / 61.6	78.8 / 56.8	85.0 / 61.9

TABLE II
COMPARISON RESULTS ON GTOT.

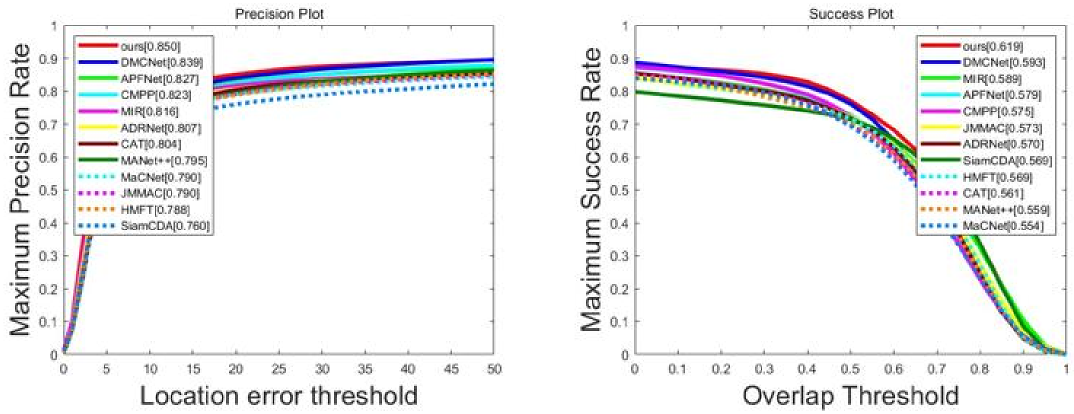
Trackers	HMFT [6]	DMCNet [5]	CMPP [8]	MTNet
PR	91.3	90.9	92.6	93.5
SR	74.9	73.3	73.8	76.0

Comparison with the SOTA Trackers

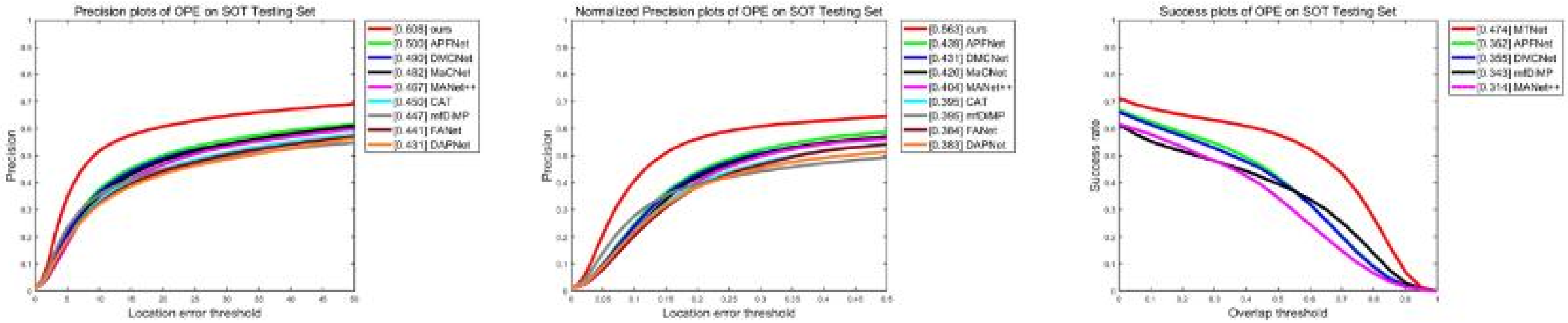
GTOT PR:0.935 SR:0.760



RGBT234 PR:0.850 SR:0.619



LasHeR PR:0.608 NPR:563 SR:0.474



Experiments

- Ablation Studies

TABLE III
ABLATION STUDY ON DIFFERENT COMPONENTS.

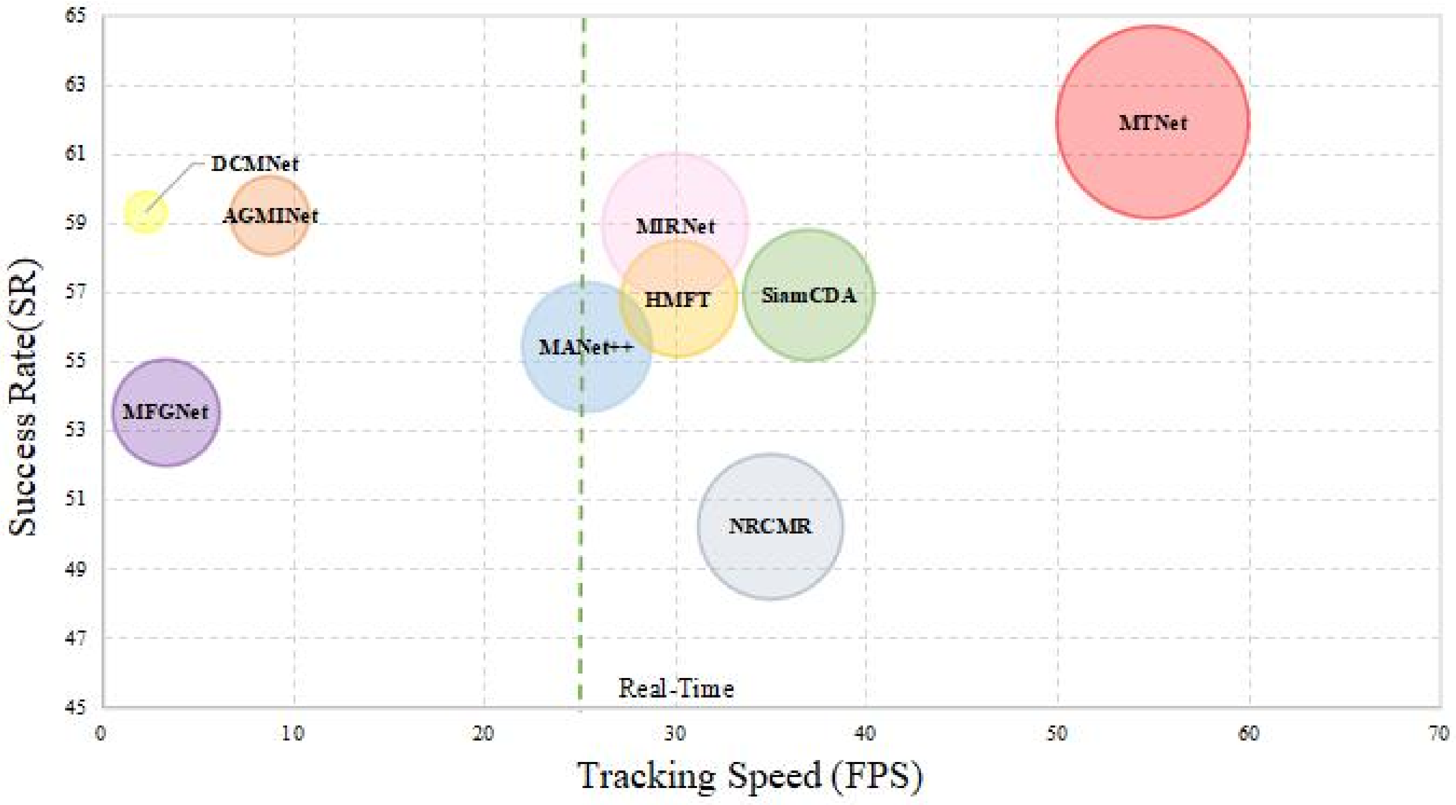
Variants	Modality-aware	Loss	Update	PR	NPR	SR
①				56.8	52.4	44.9
②	✓			58.6	54.1	46.2
③	✓	✓		59.4	55.0	46.5
④	✓	✓	✓	60.8	56.3	47.4

TABLE IV
COMPARISON OF DIFFERENT THRESHOLDS ON RGBT234.

Update interval	$N=0$	$N=2$	$N=5$
$M=60$	83.3 / 60.5	83.7 / 60.8	84.2 / 60.5
$M=70$	84.7 / 61.7	**85.0 / 61.9**	84.9 / 61.8
$M=80$	84.0 / 61.1	83.7 / 60.9	84.0 / 61.1

Experiments

- Efficiency Analysis



Experiments

- Qualitative Analysis

Conclusion

- In this work, we proposed a novel MTNet for robust RGBT tracking.
- A modality-aware network was invented to reinforce modality-specific cues from multiple perspectives, while a hybrid transformer fusion network was utilized to establish the long-distance association between the augmented features.
- The trident prediction head and the state-aware template update strategy were jointly used to a high-quality dynamic template that tackles various tracking challenges and realizes stable all weather tracking.
- Experimental results validate that our tracker achieves state-of-the-art performance on three public RGBT benchmarks while meeting real-time requirements.

IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

Thank you for your attention!

E_mail: rc_hou@smail.nju.edu.cn

https://github.com/xuboyue1999/MTNet-ICME23

南京大學 NANJING UNIVERSITY
MAGUS MediA recoGnition and UnderStanding