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

1st Ruichao Hou
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, China
rc_hou@smail.nju.edu.cn

2nd Boyue Xu
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, China
xuby@smail.nju.edu.cn

3rd Tongwei Ren*
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, China
rentw@nju.edu.cn

4th Gangshan Wu
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, China
gswu@nju.edu.cn

***Abstract*—The ability to learn robust multi-modality representation has played a critical role in the development of RGBT tracking. However, the regular fusion paradigm and the invariable tracking template remain restrictive to the feature interaction. In this paper, we propose a modality-aware tracker based on transformer, termed MTNet. Specifically, a modality-aware network is presented to explore modality-specific cues, which contains both channel aggregation and distribution module (CADM) and spatial similarity perception module (SSPM). A transformer fusion network is then applied to capturing global dependencies to reinforce instance representations. To estimate the precise location and tackle the challenges, such as scale variation and deformation, we design a trident prediction head and a dynamic update strategy which jointly maintain a reliable template for facilitating inter-frame communication. Extensive experiments validate that the proposed method achieves satisfactory results compared with the state-of-the-art competitors on three RGBT benchmarks while reaching real-time speed.**

***Index Terms*—Modality-aware, transformer, template update, RGBT tracking**

I. Introduction

RGBT tracking has been one of the emerging tasks of the computer vision community, which aims to estimate the position and scale of a pre-labeled object in a video sequence [1]. It has diverse applications in robotics, intelligent surveillance, transportation management, and unmanned vehicles [2], [3].

Recently trackers based on multi-domain learning [4] seek to enrich target expression by inserting hierarchical feature extraction [5], [6], diverse attention mechanisms [7]–[9] and attribute-aware subnetworks [10], [11]. Another type is inspired by similarity learning [12], [13], which tends to achieve fast speed. Subsequently, the latest transformer-based method [14] is proposed to push tracking performance to a new level. Nevertheless, the robust feature representation and potential inter-frame information are not explored well due to the regular fusion network and static tracking template. Some

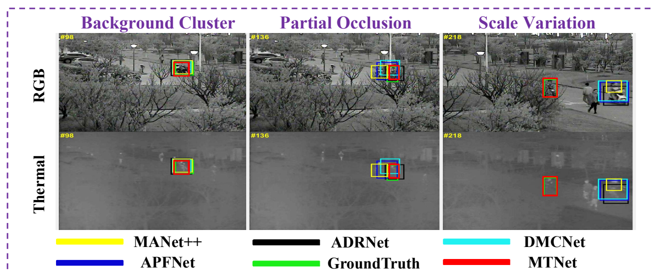

Fig. 1. Comparison results with representative trackers, *i.e.*, MANet++ [15], ADRNet [11], DMCNet [5], APFNet [10]. The MTNet performs well in complex scenarios.

visualization examples indicate most trackers still suffer from challenging factors, as shown in Fig. 1.

Building on the above analysis, we propose a novel RGBT tracker named MTNet, which addresses two issues as follows: (1) How to efficiently extract discriminative cues from heterogeneous modalities conducive to instance representation. (2) How to estimate the precise bounding box and tackle the tracking challenges. For the first issue, we design a modality-aware network to adequately emphasize meaningful features of individual patterns from multiple perspectives. It contains two cost-effective components, *i.e.*, CADM and SSPM, that make the utmost use of attention schemes for robust feature learning. Unlike existing fusion approaches [13], [14], CADM aims to produce channel-refined features and SSPM flexibly encodes spatial similarity to guide specific modal enhancement. Then, a hybrid attention-based transformer is applied to produce a correlation between the fused template and search region, which comprehensively considers global dependencies via self-attention and cross-attention. For the second issue, we define a mutual constraint loss by attaching an extra localization branch to establish the associations between the classification and regression branches for joint learning, ensuring accurate results. Instead of adopting optical flow [5], [7] or sub-networks [16] to refine the bounding box, the update

*Corresponding Author.

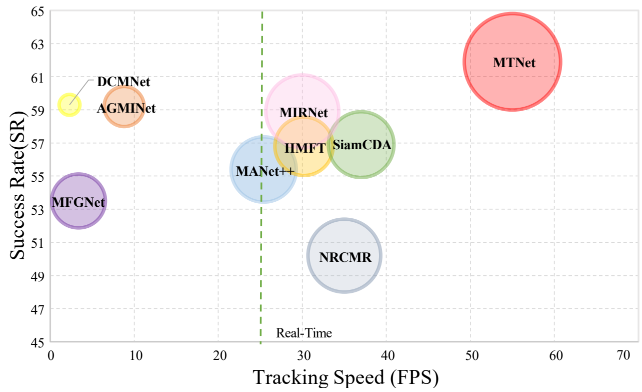

Fig. 2. Comparison with state-of-the-art trackers, *i.e.*, DMCNet [5], MIRNet [16], MANet++ [15], AGMINet [7], MFGNet [17], SiamCDA [13], NRCMR [18], HMFT [6] on RGBT234. We plot the Success Rate with respect to the Frames Per Second (FPS) tracking speed. The bubble area represents the weighted sum of the FPS and SR.

strategy attempts to maintain a reliable template for boosting inter-frame communication by accumulating confidence scores during a time interval. Experimental results prove MTNet achieves the best results and the top inference speed of 55 FPS on RGBT234, outperforming the newest trackers by a clear margin, as shown in Fig. 2.

The major contributions of this work are summarized as:

- 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.
- 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.
- Extensive experiments demonstrate that the MTNet achieves satisfactory results while running at real-time speed.

II. Related work

A. RGBT Tracking

Recently, deep learning held the dominating status of RGBT tracking, which mainly consists of two mainstream frameworks. One type of method is based on tracking by detection. For example, Lu *et al.* [5] attempted to exploit useful cues across modalities and relieve the disturbance of background clutter by proposing a duality-gated mutual condition network, which yielded competitive results. Mei *et al.* [7] presented an asymmetric global and local mutual integration network to mine heterogeneous features. Wang *et al.* [17] conducted a novel dynamic convolutional filter to fuse multi-modal features for robust tracking. To cope with multiple challenges, some variants [10], [11] developed attribute-aware sub-networks to generate modality-specific representations. Moreover, Zhang *et al.* [6] contributed an RGBT UAV tracking dataset and then proposed a baseline HMFT by combining a multi-stage fusion. Another type of work incorporates similarity learning into tracking, which aims to model the optimum matching relationship between the template and search region. For instance, Zhang *et al.* [13] designed a Siamese-based RGBT tracker with a complementary-aware multi-modal feature fusion. Feng *et al.* [14] presented a weight allocation rule to measure the reliability of shallow features, and then used the transformer to strengthen semantic information. However, there is still room for improvement in aspects of cross-modal and inter-frame cues mining.

B. Vision Transformer for Tracking

Transformer was first designed for machine translation tasks and has become the dominant structure in natural language processing. The attention mechanism is the key component in Transformer, which learns to establish dependencies between each element in the sequence [19]. Given the success of transformers in computer vision, the latest studies have applied this elegant paradigm to visual tracking. For instance, Meinhardt *et al.* [20] defined the tracking-by-attention paradigm and designed an end-to-end transformer tracker for multi-object tracking. Chen *et al.* [21] presented a transformer-based feature fusion method to replace the traditional correlation operation for building the matching relation between the template and search region. Yan *et al.* [22] proposed a novel tracker with an encoder-decoder transformer by learning spatial-temporal cues to produce satisfactory tracking results. These ground-breaking works will motivate us to bring advanced transformer architectures to RGBT tracking and promote tracking performance.

III. Methodology

A. Network Architecture

The pipeline is shown in Fig. 3. Concretely, we utilize the tailored Resnet-50 as the backbone to obtain the template and search region features. Next, the modality-aware network is invented to generate modality-specific representations. To establish the accurate matching correlation, we flatten the fused template and fused search features to vectors and then aggregate them through the hybrid transformer fusion network. Then, the prediction head with triple branches is proposed to estimate the target state. Finally, we apply the template update strategy to select the most appropriate template for refining the subsequent tracking sequences.

B. Modality-aware Network

Channel Aggregation and Distribution Module. Thermal noise or background clutter is widespread in RGBT data. Channel refinement has become an essential operation in previous works. As shown in Fig. 4, we construct a simple yet effective CADM to eliminate the redundant channels of backbone features. In the stage of channel aggregation, we first sum the template features f_R^z and f_T^z (search region features f_R^x and f_T^x) from the backbone, and then the enriched feature is embedded into the global vector d_g via the Global Average Pooling (GAP) and Fully Connected (FC) layer. The aggregation operation is defined as:

$$d_g = F_g\left(GAP\left(f_R \oplus f_T\right)\right), \tag{1}$$

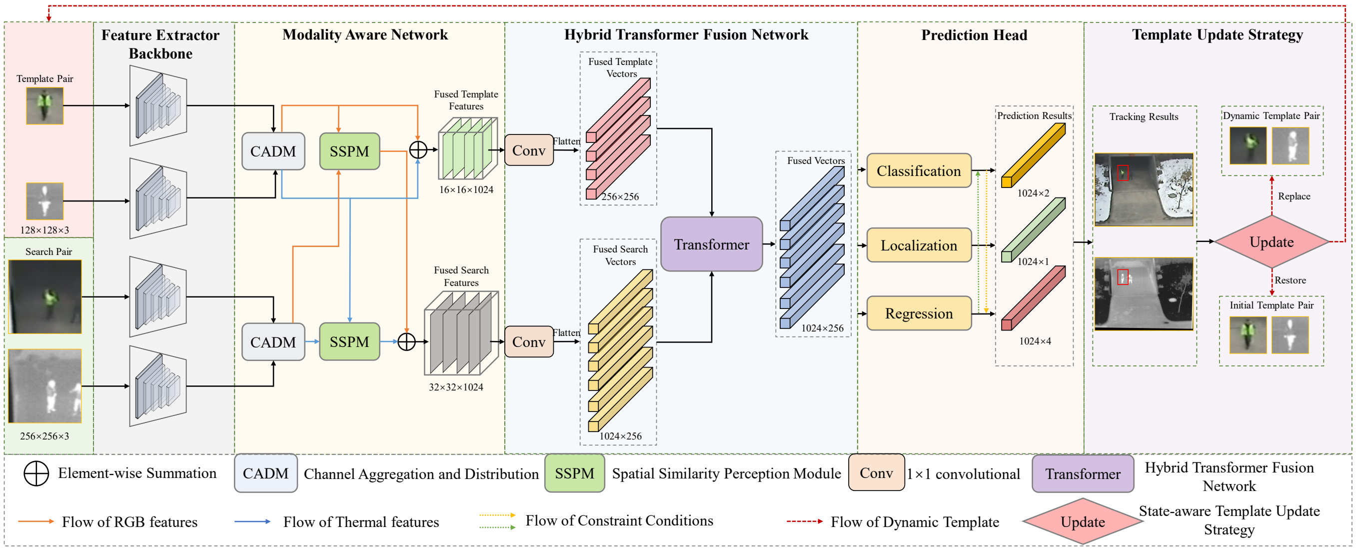

Fig. 3. The framework of MTNet, which is divided into five components, *i.e.*, Feature Extractor Backbone, Modality-aware Network, Hybrid Transformer Fusion Network, Prediction Head and Template Update Strategy.

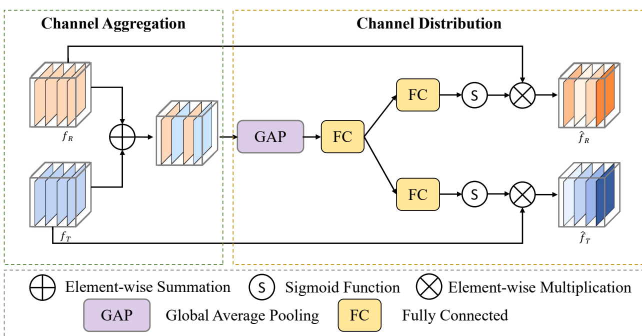

Fig. 4. Detailed design of the CADM.

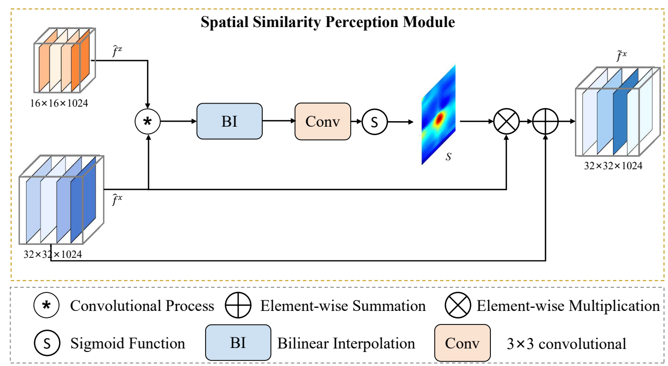

Fig. 5. Detailed design of the SSPM.

where $F(\cdot)$ is the FC layer, R and T indicate two modalities of RGB and thermal respectively. In the stage of channel distribution, we present a two-branch FC layer to get normalization channel-wise weights. Then, the channel-refined template feature maps $\hat{f}_R^z$ and $\hat{f}_T^z$ are expressed as:

$$\hat{f}_i^z = f_i^z \otimes \sigma\left(F_i\left(d_g\right)\right), i \in \{R, T\}, \tag{2}$$

where $\otimes$ is element-wise multiplication, σ is Sigmoid function. Note that CADM has the same structure but unshared weights for obtaining search region features $\hat{f}_R^x$ and $\hat{f}_T^x$.

Spatial Similarity Perception Module. SSPM depends on similarity learning to produce instance-aware residuals for further reinforcing a more reliable pattern. The diagram is shown in Fig. 5. We first take template features as instance-aware kernels to perform the convolution operation on the corresponding search region and then produce the similarity maps for the two modalities separately. The reason is that the template commonly has a higher responsive intensity in the high-quality search region, and therefore spatial similarity maps are suitable for measuring the reliability of the modality while reinforcing specific representation in the spatial domain. On account of the convolution filtering reducing the resolution of the spatial similarity map, we adopt the bilinear interpolation and convolution operation to refine the spatial similarity map S_i, which are defined as:

$$S_i = \sigma(f_{conv}(BI(\hat{f}_i^z * \hat{f}_i^x))), i \in \{R, T\}, \tag{3}$$

where $*$ denotes convolution operation, f_{conv} means 3×3 convolution operation, BI represents bilinear interpolation, σ is the Sigmoid function. The modality-aware feature maps are generated by attaching the residual connection, and we receive the joint representation by merging each modality feature map. The augmented template and search region features $\tilde{f}^z$ and $\tilde{f}^x$ can be expressed as:

$$\tilde{f}^z = \hat{f}_R^z \oplus \hat{f}_T^z, \tag{4}$$

$$\tilde{f}^x = ((\hat{f}_R^x \otimes S_R) \oplus \hat{f}_R^x) \oplus ((\hat{f}_T^x \otimes S_T) \oplus \hat{f}_T^x). \tag{5}$$

C. Hybrid Transformer Fusion Network

The powerful fusion paradigm from TransT [21] is adopted to sense the correlation between the target and search region. As shown in Fig. 6, the template feature $\tilde{f}^z$ and search feature $\tilde{f}^x$ are fed into a 1×1 convolutional layer and then reshaped to generate two vectors X^z and X^x. Then the transformer fusion network takes X^z and X^x as the input to mine meaningful features by adopting self-attention and cross-attention. The multi-head cross-attention module aims to fuse features from

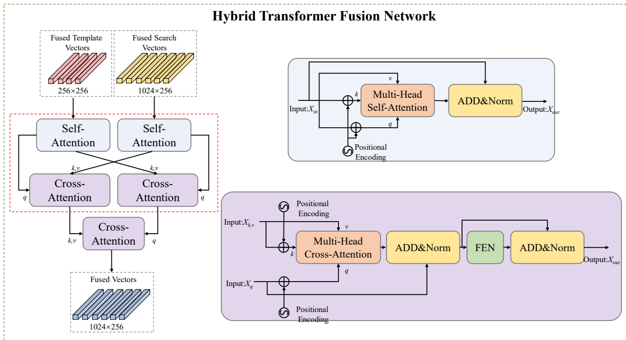

Fig. 6. Detailed design Hybrid Transformer Fusion Network.

different branches. Moreover, a feedforward network (FFN) consisting of two linear layers and a ReLU activation function boosts the fitting ability of the tracker. We build a hybrid transformer fusion network by stacking those modules four times. Finally, an extra cross-attention module is utilized to obtain the final fused vectors.

D. Trident Prediction Head

The prediction head contains three branches, *i.e.*, classification, regression and localization. Due to the prediction inconsistency between classification and regression, we insert a mutual constraint flow into binary cross-entropy loss by multiplying the normalized IoU, which aims to suppress unreasonable proposals. The classification loss is formulated as:

$$\mathcal{L}_{cls} = -\sum_{j}((y_j \log(p_j) IoU + (1-y_j)\log(1-p_j))), \quad (6)$$

where y_j defines the label of the jth sample $y_j = 1$ denotes the positive sample, p_j indicates the probability belonging to the foreground, and IoU represents the Intersection over Union between prediction and ground truth. The regression loss contains two parts: l_1-norm loss and Complete IoU loss [23], which is defined as:

$$\mathcal{L}_{reg} = \sum_{j}\Pi_{y_j=1}(\lambda_1 \mathcal{L}_1(b_j, \hat{b_j}) + \lambda_C \mathcal{L}_{CIoU}(b_j, \hat{b_j})p_j), \quad (7)$$

where b_j means the j-th bounding box, and p_j denotes the corresponding classification confidence of the positive samples. The regularization parameters λ_1 and λ_C are set to 5 and 2 respectively. Localization loss is constructed by binary cross-entropy loss and is described as:

$$\mathcal{L}_{loc} = -\sum_{j}(O_j \log(p_j^{loc}) + (1-O_j)\log(1-p_j^{loc})), \quad (8)$$

where O_j denotes the IoU scores calculated by the regression branch, and p_j^{loc} means the predicted value of the localization branch. The overall loss is defined as:

$$\mathcal{L} = n_1\mathcal{L}_{cls} + n_2\mathcal{L}_{reg} + n_3\mathcal{L}_{loc}. \quad (9)$$

where n_1, n_2 and n_3 represent the hyperparameters.

E. State-aware Template Update Strategy

In practical tracking tasks, the appearance of the target object often changes over time. 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. Specifically, the steady state is defined as the condition in which the confidence score of M consecutive frames is greater than 0.9. Once the steady state is reached, the current template will replace the initial template. If the confidence score is between 0.7 and 0.9, we reckon the tracker is in a state of transient steady and the template remains constant during this interval. If the confidence is lower than 0.7 and has accumulated up to N times, the tracker may struggle in an unstable state, and the current template is restored by an initial template. To pursue the best performance, we set different update intervals for each state.

IV. Experiments

A. Datasets and Metrics

In this paper, we conduct comparative experiments with high-performance competitors on three popular RGBT benchmarks, *i.e.*, GTOT [24], RGBT234 [1] and LasHeR [2]. Following mainstream works, we employ two classical metrics, Precision Rate (PR) and Success Rate (SR) to measure tracking performance. For a fair comparison, we set the threshold of GTOT to 5 pixels and RGBT234/LasHeR to 20 pixels considering the inherent inconsistent image resolution between different datasets. Moreover, we apply the Normalized Precision Rate (NPR) metric to alleviate the influence of the resolution for testing the LasHeR.

B. Implementation Details

The MTNet is implemented on the PyTorch 1.10 platform with two NVIDIA RTX3090 GPUs with 24GB memory. In the offline training phase, MTNet is trained on the LasHeR. The AdamW optimizer is utilized to update the model, and the initial learning rate and weight decay are both set to $1e^{-4}$. The model is trained over 40 epochs, each containing 1,000 iterations, with a batch size of 16. After the first 20 epochs, the learning rate is decreased by a factor of 10. Hyperparameters n_1, n_2, and n_3 are set to 8, 5, and 1, respectively. During the online tracking phase, the prediction head generated 1,024 proposals, which are ranked based on window penalty and location logits. To test the GTOT, update intervals are set to $\{50, 2\}$ due to the small scale of the dataset. For other datasets, update intervals are set to $\{70, 2\}$. The best tracking result is determined by selecting the bounding box with the highest confidence score.

C. Comparison with the State-of-the-Art

The proposed tracker is compared with 11 latest methods, *i.e.*, DMCNet [5], MIRNet [16], APFNet [10], AGMINet [7], MFGNet [17], SiamCDA [13], RMWT [14], HMFT [6], CMPP [8], MANet++ [15], ADRNet [11].

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

Overall Performance. As reported in Table I, MTNet outperforms all other competitors with 85.0%/61.9% in PR/SR, achieving performance gains of 1.0%/2.7% over the second-ranked tracker AGMINet on RGBT234. Besides, MTNet achieves the best results on GTOT, with PR/SR reaching 93.5%/76%, as given in Table II. Specifically, MTNet outperforms the HMFT by 2.2%/1.1% in PR/SR. Fig. 7 shows that MTNet obtains the best ranking on LasHeR, with 60.9%, 56.3% and 47.4% in PR, NPR and SR. Furthermore, even when compared to the retrained tracker mfDiMP [25] for comparison, MTNet still surpasses it by 0.9%/0.7%.

Attribute-based Performance. The attribute-based comparison on RGBT234 are presented in Table I. The attributes consist of no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM), low resolution (LR), thermal crossover (TC) and background cluster (BC). Experimental results suggest that the proposed approach works well in adverse conditions. Specifically, compared to the transformer-based method RMWT, MTNet achieves performance gains of 3.4%/0.8%, 4.7%/2%, 2.9%/0.2%, 6.6%/3.8% and 2.9%/0.7% in the attributes of HO, DEF, SV, MB, and CM, respectively. One of the important reasons for the superior performance of MTNet is the incorporation of CADM and SSPM modules, which enable the learning of robust multi-modal representations. Additionally, the proposed state-aware template update strategy helps to mitigate the impact of unreliable appearance features. In the attributes of NO, LI, LR, and BC, tracking-by-detection-based methods, *i.e.*, MIRNet, AGMINet, and DMCNet achieve the best performance due to their online training mechanism and extra refinement network, but at the cost of increased complexity. To summarize, MTNet strikes a good balance between efficiency and performance compared to other competing trackers.

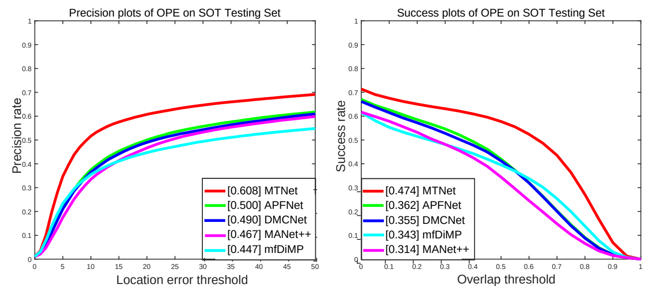

Fig. 7. The evaluation curve on LasHeR.

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

D. Ablation Study

Components Analysis. To further validate the feasibility of each contribution, we implement three variants and test them on LasHeR datasets, *i.e.*, ① is a base model, which is by TransT [21] and integrates dual-pattern features via simple element addition. ② incorporates the modality-aware network into the baseline tracker. ③ combines prediction head with mutual constraint loss on the basis of ②. ④ is the final version equipped with the template update strategy.

According to the tracking results reported in Table III, we can draw the following conclusions: 1) The modality-aware network flexibly exploits cues between dual patterns to enhance modality-aware representation. 2) The trident prediction head improves localization accuracy by unifying the distribution between each branch. 3) The template update strategy introduces additional temporal context to alleviate the appearance variation issue, which boosts overall performance.

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

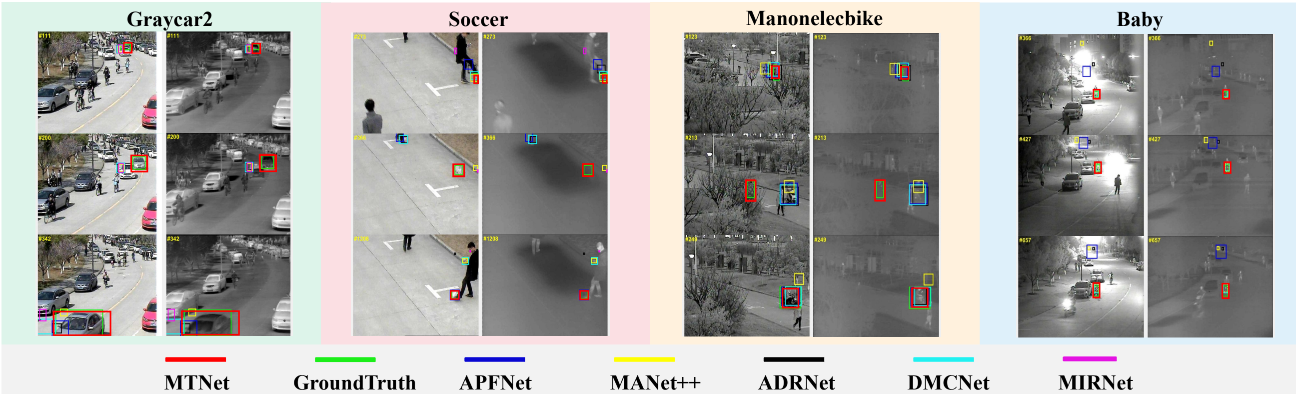

Fig. 8. Qualitative comparison between MTNet and other trackers on four challenging sequences.

Parameters Analysis. To measure the impact on performance, we set $M = \{60, 70, 80\}$, $N = \{0, 2, 5\}$ to carry out evaluation and the comparison results on RGBT234 are reported in Table IV. We observe the best metrics are determined by the parameters $\{70, 2\}$. When the interval is longer or shorter it may lead to a suboptimal template. In addition, instantaneously resetting the current template may cause misjudgment of the state. In both cases, the best results may not be achieved. Hence, selecting an appropriate update interval can effectively improve tracking performance.

Qualitative Analysis. The qualitative comparison is shown in Fig. 8. Thanks to the modality-aware representation and reliable template, the proposed tracker performs well when encountering multiple challenges, especially FM, HO, SV, and MB. Therefore, the superiority of MTNet has been adequately verified again via intuitive qualitative comparison.

V. 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. Experiments verify that the proposed method attains impressive performance compared to state-of-the-art trackers while achieving real-time requirements.

Acknowledgment

This work is supported by the National Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the program B for Outstanding Ph.D. candidate of Nanjing University, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

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, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, “Lasher: A large-scale high-diversity benchmark for rgbt tracking,” *IEEE TIP*, 2021.
[3] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, “VIF-Net: An unsupervised framework for infrared and visible image fusion,” *IEEE TCI*, 2020.
[4] I. Jung, J. Son, M. Baek, and B. Han, “Real-time mdnet,” in *ECCV*, 2018.
[5] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, “Duality-gated mutual condition network for rgbt tracking,” *IEEE TNNLS*, 2022.
[6] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, “Visible-thermal uav tracking: A large-scale benchmark and new baseline,” in *CVPR*, 2022.
[7] J. Mei, Y. Liu, C. Wang, D. Zhou, R. Nie, and J. Cao, “Asymmetric global–local mutual integration network for rgbt tracking,” *IEEE TIM*, 2022.
[8] 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.
[9] W. Xia, D. Zhou, J. Cao, Y. Liu, and R. Hou, “Cirnet: An improved rgbt tracking via cross-modality interaction and re-identification,” *Neurocomputing*, 2022.
[10] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, “Attribute-based progressive fusion network for rgbt tracking,” in *AAAI*, 2022.
[11] P. Zhang, D. Wang, H. Lu, and X. Yang, “Learning adaptive attribute-driven representation for real-time rgb-t tracking,” *IJCV*, 2021.
[12] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, and P. Torr, “Fully-convolutional siamese networks for object tracking,” in *ECCV Workshops*, 2016.
[13] T. Zhang, X. Liu, Q. Zhang, and J. Han, “Siamcda: Complementarity-and distractor-aware rgb-t tracking based on siamese network,” *IEEE TCSVT*, 2021.
[14] M. Feng and J. Su, “Learning reliable modal weight with transformer for robust rgbt tracking,” *KBS*, 2022.
[15] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, “Rgbt tracking via multi-adapter network with hierarchical divergence loss,” *IEEE TIP*, 2021.
[16] R. Hou, T. Ren, and G. Wu, “Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement,” in *ICME*, 2022.
[17] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, “Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking,” *IEEE TMM*, 2022.
[18] C. Li, Z. Xiang, J. Tang, B. Luo, and F. Wang, “Rgbt tracking via noise-robust cross-modal ranking,” *IEEE TNNLS*, 2021.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *NeurIPS*, 2017.
[20] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, “Trackformer: Multi-object tracking with transformers,” in *CVPR*, 2022.
[21] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, “Transformer tracking,” in *CVPR*, 2021.
[22] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, “Learning spatio-temporal transformer for visual tracking,” in *ICCV*, 2021.
[23] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-iou loss: Faster and better learning for bounding box regression,” in *AAAI*, 2020.
[24] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, “Learning collaborative sparse representation for grayscale-thermal tracking,” *IEEE TIP*, 2016.
[25] 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 *ICCV Workshops*, 2019.