2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Probability-Driven Gating for Resilient Multi-Modal Tracking in Robotic Systems

Huan Wang[1], Haomin Chen[1], Pengcheng Du[1], Pengju Si[1], Baofeng Ji[1] and Yongming Yang[2]

***Abstract*— The deployment of robots in unstructured environments demands perception systems that are both accurate and resilient. While RGB-Thermal (RGB-T) fusion is promising, current trackers often fail due to rigid, non-adaptive fusion strategies and underutilized cross-modal cues, compromising reliability for robotics. We introduce DSTrack, a novel tracking framework that embeds two core mechanisms for robotic robustness: a Probability-Gated Dynamic Switch and a Synergistic Multi-Domain Enhancement Network. The switch acts as an online decision-maker, allowing the robot to dynamically select the most reliable fusion path based on real-time confidence estimation, enabling crucial adaptation to scene changes. The enhancement network concurrently strengthens target representations within each modality through tri-domain (channel, spatial, frequency) refinement and establishes compensatory links between modalities via a cross-attention module, ensuring performance even during partial sensor degradation. Extensive evaluations on RGB-T benchmarks demonstrate state-of-the-art accuracy. More critically, DSTrack exhibits key properties for robotic integration: real-time environmental adaptability, inherent sensor fault tolerance, and consistent output for downstream planning.**

I. INTRODUCTION

RGB-T tracking [1] has emerged as a pivotal research frontier in visual object tracking, witnessing remarkable advancements in recent years due to its critical applications in video surveillance systems [2], crowd counting [3], and pedestrian tracking [4]. RGB-T tracking synergistically combines single object tracking (SOT) with multi-modal perception by simultaneously processing visible (RGB) and thermal infrared (TIR) modalities. The framework initializes with dual-modality appearance templates and progressively estimates target positions and scales in subsequent frames. Capitalizing on RGB's rich spectral information and TIR's environmental robustness against illumination variations, this bimodal approach ensures reliable tracking performance under challenging conditions, including low-light scenarios and meteorological disturbances.

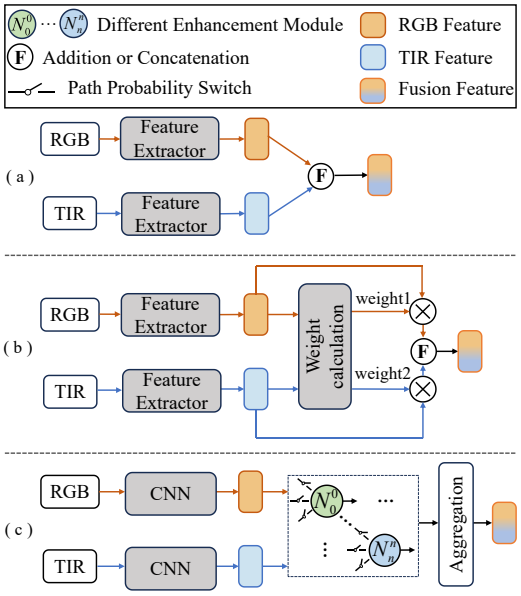

Fig. 1. Comparison of our model with other previous models. (a) and (b) denote the representation learning model and the feature fusion model. (c) The of proposed DSTrack model.

Existing RGB-T tracking methods predominantly fall into two categories, as illustrated in Fig. 1(a-b). The first category emphasizes representation learning for modality feature enhancement. Representative works include [5], which develop dual-stream adapters to extract modality-shared and modality-specific features through parallel convolutional pathways. Building on frequency domain analysis, Li et al. [6] propose a frequency-mixed perception module that strategically combines high-frequency components from individual modalities with low-frequency in cross-modal interactions to amplify discriminative features. While these approaches demonstrate progress in intra-modal feature learning, they exhibit critical limitations in cross-modal fusion, since most implementations resort to elementary operations such as linear superposition or channel concatenation Fig. 1(a). Such simplistic fusion paradigms fundamentally fail to capture nonlinear feature interdependencies between modalities, resulting in suboptimal exploitation of cross-modal complementary relationships.

The second paradigm focuses on cross-modal fusion ar-

This work was supported in part by National Natural Science Foundation of China (62571181, 62201200), China Postdoctoral Science Foundation (2025M781634, 2025M783507), the Program for Science & Technology Innovation Talents in Universities of Henan Province (26HASTIT011), the Scientific and Technological Project of Henan Province (252102211046, 262102211006), the Aeronautical Science Foundation of China (20240001042002), the Key Scientific Research Projects of Universities in Henan Province (25A120005), the Training Program for Young Backbone Teachers in Higher Education Institutions of Henan Province (2025GGJS040), the National Key Research and Development Program of China (2024YFB2907700), the Key Technological Breakthrough Projects for Major Industries in Henan Province (251000220100, 251000220300) and the Major Science and Technology Projects of Longmen Laboratory (231100220200, 231100220300, 231100220400).(Corresponding author: Yongming Yang.)

[1]Huan Wang, Haomin Chen, Pengcheng Du, Pengju Si and Baofeng Ji are with School of Information Engineering, Henan University of Science and Technology, Luoyang, 471023, China, and also with Longmen Laboratory, Luoyang, 471023, China

[2]Yongming Yang. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110169, China yangyongming@sia.cn

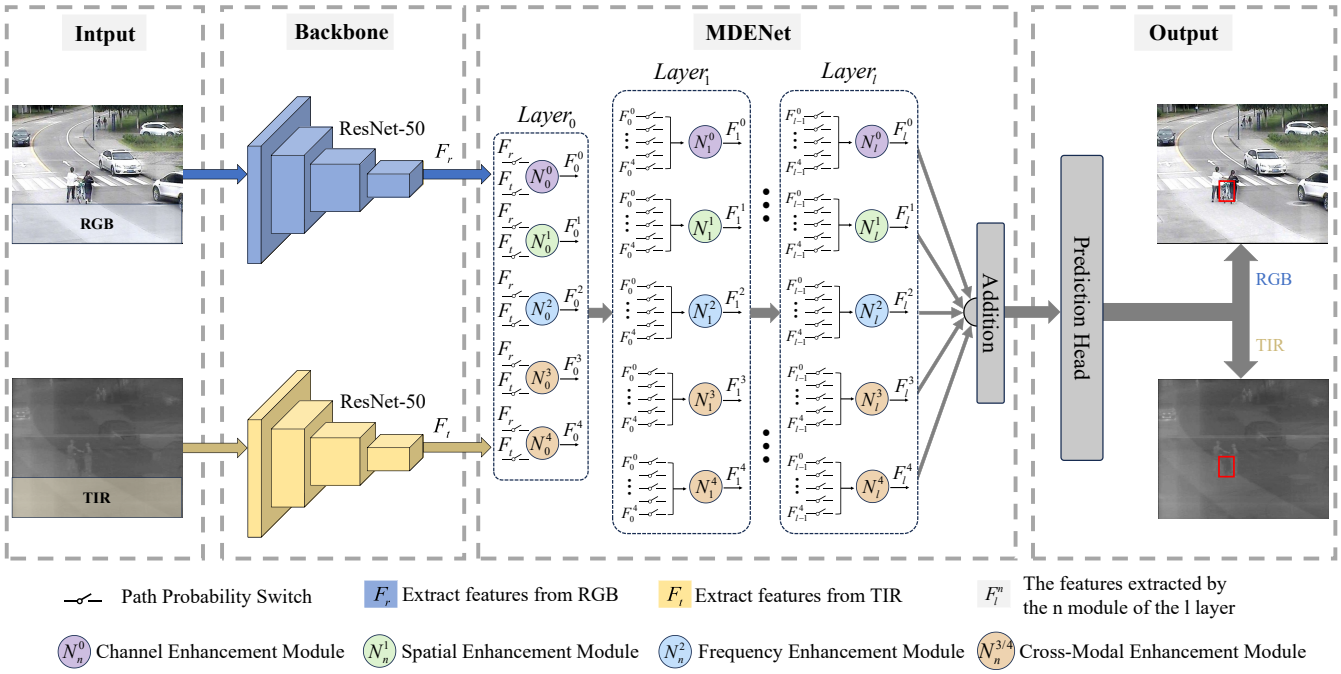

Fig. 2. The overall framework of our method. Our backbone is a bilateral ResNet-50 with shared weights. The proposed MDENet, which is composed of five distinct enhancement modules, can effectively enhance and fuse information from different modalities. It is embedded after ResNet-50. Finally, we add up the features processed by MDENet and feed them into the prediction head to obtain the final tracking result.

chitecture design, as depicted in Fig. 1(b). Representative works leverage hybrid attention mechanisms [7] to establish inter-branch feature interactions, where cross-attention layers selectively enhance primary modality features while suppressing noise from degraded modalities. Building on modal quality assessment, Liu et al. [8] develop dynamic weighting networks that predict modality reliability indices to guide feature fusion. While these approaches demonstrate improved cross-modal interaction through learnable weighting strategies, the modules connectivity patterns and weight generation mechanisms remain fixed during the inference phase, which fundamentally constrains adaptability to dynamic environmental conditions. Consequently, existing fusion architectures inevitably compromise tracking optimality when confronted with complex real-world scenarios requiring structural reconfiguration.

To address these issues, we propose a novel dynamic enhancement network with switch for RGB-T tracking (DSTrack). As shown in Fig. 1 (c), it illustrates the main differences between our method and existing methods. The switch first extracts key information from each modality through a pooling operation, and then performs a series of steps to generate a final decision, which determines whether to select the current enhancement module, thereby dynamically deciding the combination form. This enables DSTrack to flexibly adjust the network structure based on features such as modality quality and scene complexity of the input data, ensuring optimal fusion of multi-modal features.

Our main contributions are summarized as follows:

- We propose a novel RGB-T tracking framework that integrates switch-guided dynamic fusion with multi-domain enhancement. The architecture enables flexible network structure reconfiguration through learnable gating mechanisms.
- We propose a multi-domain enhancement network for progressive feature enhancement. It constructs tri-domain intra-modal enhancements (channel-spatial-frequency) via mixed attention and wavelet decomposition, along with two directions cross-modal compensation using attention bridges to support progressive feature refinement.
- We conduct extensive experiments on four RGB-T tracking benchmark datasets, demonstrating that our method achieves state-of-the-art performance. Specifically, our tracker achieves PR/SR scores of 89.3%/65.5% and 85.5%/72.7% on the RGBT234 and VTUAV datasets, respectively.

II. Method

A. Multi-Domain Enhancement Network

To fully leverage the information from RGB and TIR modalities in different domains, we design a Multi-Domain Enhancement Network composed of intra-modality enhancement modules and inter-modality enhancement modules, aiming to improve the discriminability of modalities and cross-modal collaboration capabilities. For the intra-modal enhancement module, we design the channel enhancement module (CEM), the spatial enhancement module (SEM), and the frequency enhancement module (FEM) to optimize the

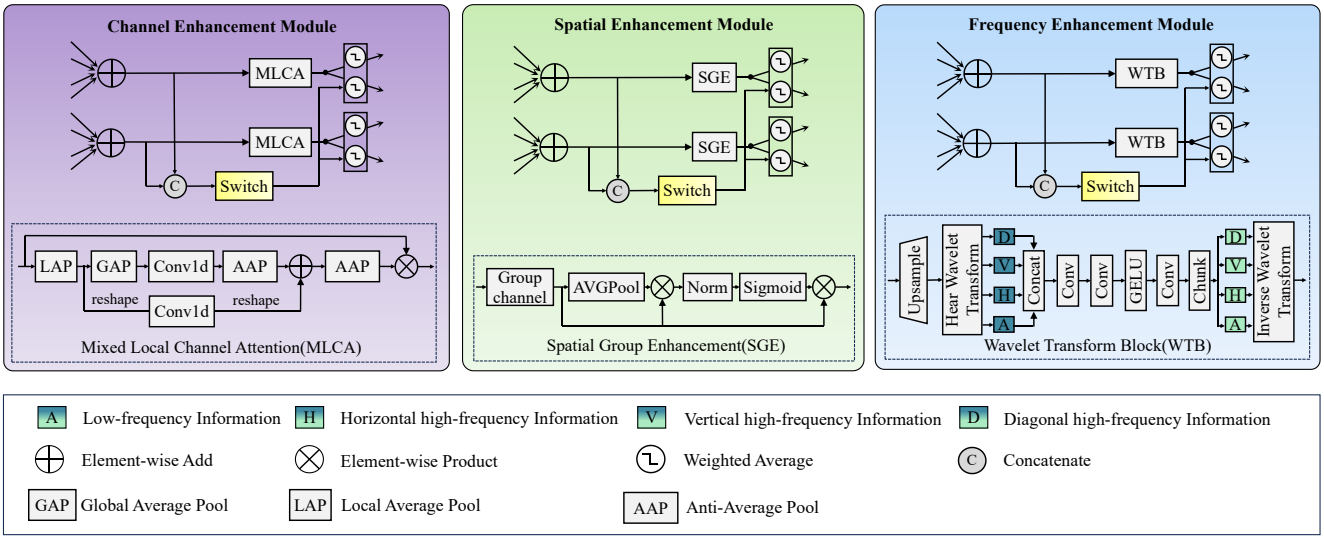

Fig. 3. Illustration of the intra-modality enhancement module. It includes mixed local channel attention module (MLCA), spatial group enhancement module (SGE), wavelet transform module (WTB), and combined with switch to form channel, spatial, and frequency enhancement modules.

feature expression from channel relationship, spatial distribution, and frequency characteristics, respectively. For the inter-modal enhancement module, we introduce the RGB to TIR cross-modal enhancement module ($\mathrm{CMEM}_{r\to t}$) and the TIR to RGB cross-modal enhancement module ($\mathrm{CMEM}_{t\to r}$). We stack the above five modules to gradually optimize feature representation within the modality through dense connections, and enhance the information transmission between RGB and TIR through multi-layer interaction. This design allows the network to effectively integrate the advantages of both modalities, achieving complementary feature expressions through multi-domain enhancement, thereby improving the stability and accuracy of target tracking.

1) Intra-Modality Enhancement Module: In multi-modal feature modeling, adequate extraction of intra-modal information is essential for achieving cross-modal collaboration. To this end, this section conducts in-depth modeling of RGB and thermal infrared (TIR) features from three perspectives: channel domain, spatial domain, and frequency domain, aiming to enhance the feature representation capabilities of the dual modalities. The intra-modality enhancement module is shown in Fig. 3.

Channel Enhancement Module. To overcome limitations of channel-only attention, we propose a Mixed Local Channel Attention (MLCA) module that incorporates local spatial information. The input feature map is divided into local blocks, which are vectorized to capture regional details. Two parallel branches extract global context (via global average pooling) and local features (via local pooling and 1D convolution). Features are then restored via anti-average pooling and fused to produce the final channel attention map.

Spatial Enhancement Module. To handle challenges like occlusion and deformation, we introduce a Spatial Group Attention (SGE) module. Input features are grouped channel-wise, and each group is modulated by a attention mask derived from global average pooling. Two learnable parameters (w, b) scale and shift the attention, which is applied via Sigmoid and element-wise multiplication to refine spatial feature distributions.

Frequency Enhancement Module. We propose a Wavelet Transform Block (WTB) to enhance frequency-domain details. The input is upsampled and decomposed via wavelet transform into four sub-bands: low-frequency (A), horizontal (H), vertical (V), and diagonal (D) high-frequencies. These are concatenated, processed by convolutions and GELU, then split and reconstructed via inverse wavelet transform to recover spatial features.

Inter-Modal Enhancement Module. We introduce a bidirectional cross-attention mechanism for RGB-T interaction. For example, in RGBTIR enhancement, TIR features serve as query Q_{tir}, while RGB features provide key K_{rgb} and value V_{rgb}. After normalization and 1×1 convolution, attention scores are computed via dot product, softmax-normalized, and used to aggregate V_{rgb}. The result is added to the original TIR feature for enhancement.

B. Dynamic Fusion Structure

Each module in our multi-domain enhancement network is treated as a node, and nodes in different layers are connected to each other in a fully connected manner. In the first layer, the features F_{rgb} and F_{tir} are extracted separately and passed as inputs to the nodes. Nodes in subsequent layers receive features from all nodes in the previous layer in a fully connected manner, achieving richer feature interaction and transmission. The input of each node can be represented as follows:

$$O_i^{(l)} = \begin{cases} \sum_{n=0}^{N-1} R_{n,i}^{(l-1)} \cdot O_n^{(l-1)} & if\ l > 1, \\ (F_{rgb}, F_{tir}) & if\ l = 1. \end{cases} \tag{1}$$

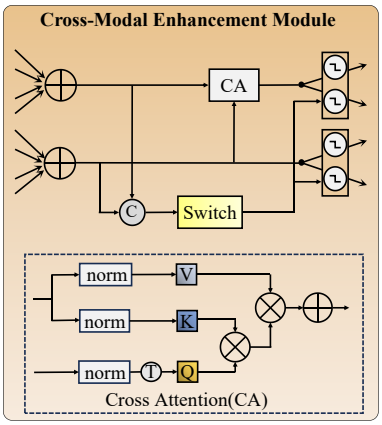

Fig. 4. Illustration of the inter-modal enhancement module. The module generates specific enhanced modality features based on query. When TIR features are used as the query, it produces enhanced TIR modality features. Conversely, when RGB features serve as the query, it outputs enhanced RGB modality features.

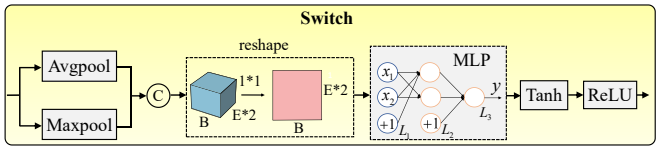

Fig. 5. Illustration of Switch. It can indicate whether the module is combined with other enhancement modules in the next layer based on the input dual-modal features.

where N denotes the number of nodes in each layer, and $O_n^{(l-1)}$ represents the output of the n-th node in the $l-1$ layer. $R_{n,i}^{(l-1)}$ indicates the backward transmission probability of the output feature from the n-th node in the $l-1$ layer, ranging between 0 and 1. This probability is determined by the switch. In tracking scenarios where modal quality is severely unbalanced, the mutual enhancement between modalities might bring feature contamination. Since not all enhancement modules are always necessary in different tracking scenarios, we introduce a dynamic switch mechanism and combine it with a multi-domain enhancement network to form a dynamic fusion structure. The design of the switch module is shown in Fig. 5.

In multi-domain enhancement network, each enhancement module is embedded with a switch to determine whether the module should be activated and passed to the next layer. Specifically, the switch extracts global statistical information from input features via global average pooling and global max pooling. Then, these features are fed into MLP to learn the interaction relationships and fusion strategies between different modalities. Subsequently, the output of MLP is activated by the $Tanh$ and $ReLU$ functions to ensure the non-linear expression ability of the routing decisions, making the selection of different enhancement modules more flexible.

TABLE I

THE PR, NPR, AND SR SCORES (%) OF VARIOUS TRACKERS ON TWO DATASETS. HIGHER VALUE INDICATE BETTER PERFORMANCE. THE BEST AND SECOND RESULTS ARE DISPLAYED IN RED AND BLUE FONTS.

Methods	RGBT234		LasHeR			FPS
	PR	SR	PR	NPR	SR	
DRGCNet [13]	82.5	58.1	48.3	42.3	33.8	4.9
MACFT [7]	85.7	62.2	65.3	-	51.4	31.7
CMD [14]	82.4	58.4	59.0	54.6	46.4	30
ViPT [15]	83.5	61.7	65.1	-	52.5	-
TBSI [16]	87.1	63.7	69.2	65.7	55.6	36.2
STTANet [17]	85.5	63.2	66.7	-	53.4	18.6
CAT++ [18]	84.0	59.2	50.9	44.4	35.6	14
OneTracker [19]	85.7	64.2	67.2	-	53.8	-
UnTrack [20]	84.2	62.5	66.7	-	53.6	-
SDSTrack [21]	84.8	62.5	66.5	-	53.1	20.9
DSTrack	89.3	65.5	69.3	64.8	54.3	28.6

TABLE II

COMPARISON WITH STATE-OF-THE-ART TRACKERS ON RGBT210 [9] DATASET.

Methods	Publish	Backbone	RGBT210	
			PR	SR
mfDiMP [22]	ICCVW 2019	ResNet-50	78.6	55.5
CAT [23]	ECCV 2020	VGG-M	79.2	53.3
APFNet [24]	AAAI 2022	VGG-M	79.9	54.9
DMCNet [25]	TNNLS 2022	VGG-M	79.7	55.5
HMFT [12]	CVPR 2022	ResNet-50	78.6	53.5
MFG [26]	TMM 2022	ResNet-18	74.9	46.7
QAT [8]	ACM MM 2023	ResNet-50	86.8	61.9
TBSI [16]	CVPR 2023	ViT-B	85.3	62.5
ViPT [15]	CVPR 2023	ViT-B	83.5	61.7
CAT++ [18]	TIP 2024	VGG-M	82.2	56.1
STTANet [17]	IEEE TIM 2024	ViT-B	82.5	60.2
DSTrack	-	ResNet-50	87.2	62.9

III. Experiments

A. Dataset and Evaluation Metrics

To comprehensively evaluate the performance of DSTrack, we select four publicly available datasets widely used in the RGB-T tracking field for experiments, namely RGBT210 [9], RGBT234 [10], LasHeR [11], and VTUAV [12]. These datasets cover different scenarios and challenges, which can fully test the performance of RGB-T tracking methods in various situations.

Following previous research conventions, Precision Rate (PR), Success Rate (SR), and Normalized Precision Rate (NPR) are utilized as evaluation metrics in the field of RGB-T tracking. PR is utilized to measure the proportion of frames where the distance between the predicted bounding box center and the true target center is within a specific threshold (usually 20 pixels), reflecting the accuracy of the model in locating the target. SR calculates the percentage of frames where the Intersection over Union (IoU) between the predicted bounding boxes and the true bounding boxes is greater than a given threshold, reflecting the ability of the method to estimate the target scale. NPR normalizes PR to eliminate the impact of image resolution and target size on evaluation.

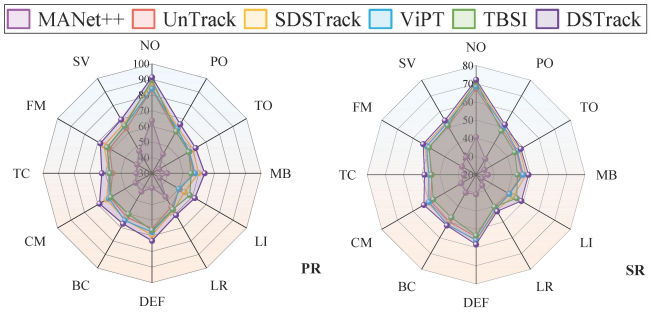

Fig. 6. The SR and PR scores (%) of MANet++, UnTrack, SDSTrack, ViPT, TBSI and DSTrack under different challenging attributes on the LasHeR dataset.

B. Implementation Details

We implement DSTrack model using PyTorch [27] framework and conduct experiments on a workstation equipped with NVIDIA RTX 3090 GPU and 24GB memory. Our model adopts ToMP [28] as the base tracker, which utilizes the first four convolutional blocks of ResNet-50 [29] as the feature extractor. We utilize the pre-trained model provided by ToMP50 [28] as the initialization parameter for the feature extractor, while the remaining parameters in our network model are randomly initialized. During training, we set the batch size to 8 and employ the AdamW optimizer [30] with a weight decay coefficient of 1×10^{-4} and an initial learning rate of 2×10^{-4}. The DSTrack model utilizes stochastic gradient descent (SGD) to minimize the classification and regression loss functions, with learning rates of 1×10^{-5} for the backbone network and prediction head. The learning rate of multi-domain enhancement network is set to 2×10^{-6}, which is trained for 100 epochs. For dataset configuration, we train the model on the LasHeR training set and use the weight files trained on LasHeR to evaluate RGBT210, RGBT234, and LasHeR testing sets. When evaluating the VTUAV dataset, we specifically utilize its own training set for model training to provide support for the evaluation of the dataset.

C. Comparison with State-of-the-art Methods

To comprehensively evaluate the performance of our method, we compare our method with previous state-of-the-art RGB-T tracking methods on four benchmarks, including RGBT210 [9], RGBT234 [10], LasHeR [11] and VTUAV [12]. The comprehensive comparison results are presented in Table I, II, III, which demonstrates the excellent performance of our method in various key indicators, validating its effectiveness in RGB-T tracking tasks.

D. Ablation Studies

In this section, we conduct a series of ablation studies on the LasHeR [11] and RGBT234 [10] datasets to evaluate the effectiveness of the proposed individual components.

1) Analysis of Multi-domain Enhancement Network : To investigate the optimal number of layers of the multi-domain enhancement network (MDENet), we conduct experiments on MDENet with different layers number. The performance of 2, 3, and 4 layers MDENet is shown in Table IV. The experimental results show that when the number of layers is set to 2, the model performs best on both datasets.

2) Analysis of Switch: To investigate the impact of switch, we remove it and the entire model becomes fixed fusion structure. The experimental results are shown in Table IV. The absence of switch significantly impairs model performance. When the number of layers is set to 2, on the LasHeR dataset, PR and SR decrease by 2.2% and 1.6% respectively; On the RGBT234 dataset, PR and SR decrease by 3.2% and 2.4% respectively. This fully verifies the importance of switch that can adaptively adjust the fusion structure, which is crucial for improving model performance.

3) Analysis of Enhancement Modules: We evaluate the effectiveness of each enhancement module in the multi-domain enhancement network. The experimental results are shown in Table V. They show that the absence of each enhancement module leads to a significant decrease in model performance. Among them, the removal of RGB-to-TIR enhancement module (N3) shows the most obvious performance decline, and PR and SR decrease by 2.1%/1.2% and 2.7%/1.6% on the RGBT234 and LasHeR datasets. This experiment indicates that each of our enhancement modules plays an

TABLE III
COMPARISON WITH STATE-OF-THE-ART TRACKERS ON VTUAV [12] DATASET.

Methods	Publish	Backbone	VTUAV	
			PR	SR
DAFNet [31]	ICCVW 2019	VGG-M	62.0	45.8
FSRPN [32]	ICCVW 2019	ResNet-50	65.3	54.4
mfDiMP [22]	ICCVW 2019	ResNet-50	67.3	55.4
ADRNet [33]	IJCV 2021	VGG-M	62.2	46.6
TransT [34]	CVPR 2021	ResNet-50	74.4	63.6
HMFT [12]	CVPR 2022	ResNet-50	75.8	62.7
MACFT [7]	Sensors 2023	ViT-B	80.1	66.8
DSTrack	-	ResNet-50	85.5	72.7

TABLE IV
ABLATION STUDIES OF LAYERS AND SWITCHS IN MDENET.

Layers	Switch	RGBT234		LasHeR		DSTrack
		PR	SR	PR	SR	Params
2	√	**89.3**	**65.5**	**69.3**	**54.3**	19.5M
3	√	87.0	63.8	68.9	54.1	29.5M
4	√	87.1	63.9	68.0	53.8	37.1M
1	×	85.9	63.0	67.0	52.7	7.2M
2	×	86.1	63.1	67.1	52.7	14.3M
3	×	86.5	63.3	67.2	52.9	21.5M

TABLE V
ABLATION OF THE DIFFERENT ENHANCEMENT MODULES IN MDENET.

N0	N1	N2	N3	N4	RGBT234		LasHeR		DSTrack
					PR	SR	PR	SR	Params
×	√	√	√	√	87.0	63.9	68.3	53.7	18.5M
√	×	√	√	√	87.5	64.4	67.0	52.9	18.5M
√	√	×	√	√	87.9	64.7	68.1	53.6	5.3M
√	√	√	×	√	87.2	64.3	66.6	52.7	18.0M
√	√	√	√	×	88.3	64.8	66.2	52.2	18.0M
√	√	√	√	√	**89.3**	**65.5**	**69.3**	**54.3**	19.5M

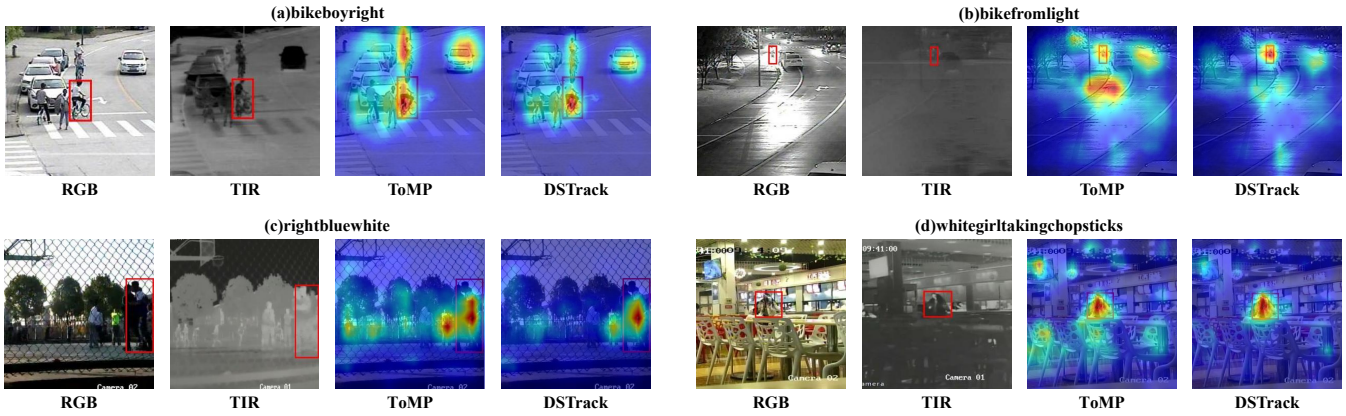

Fig. 7. Visualization of heatmaps of DSTrack and ToMP [28] on four sequences from the LasHeR [11] dataset.

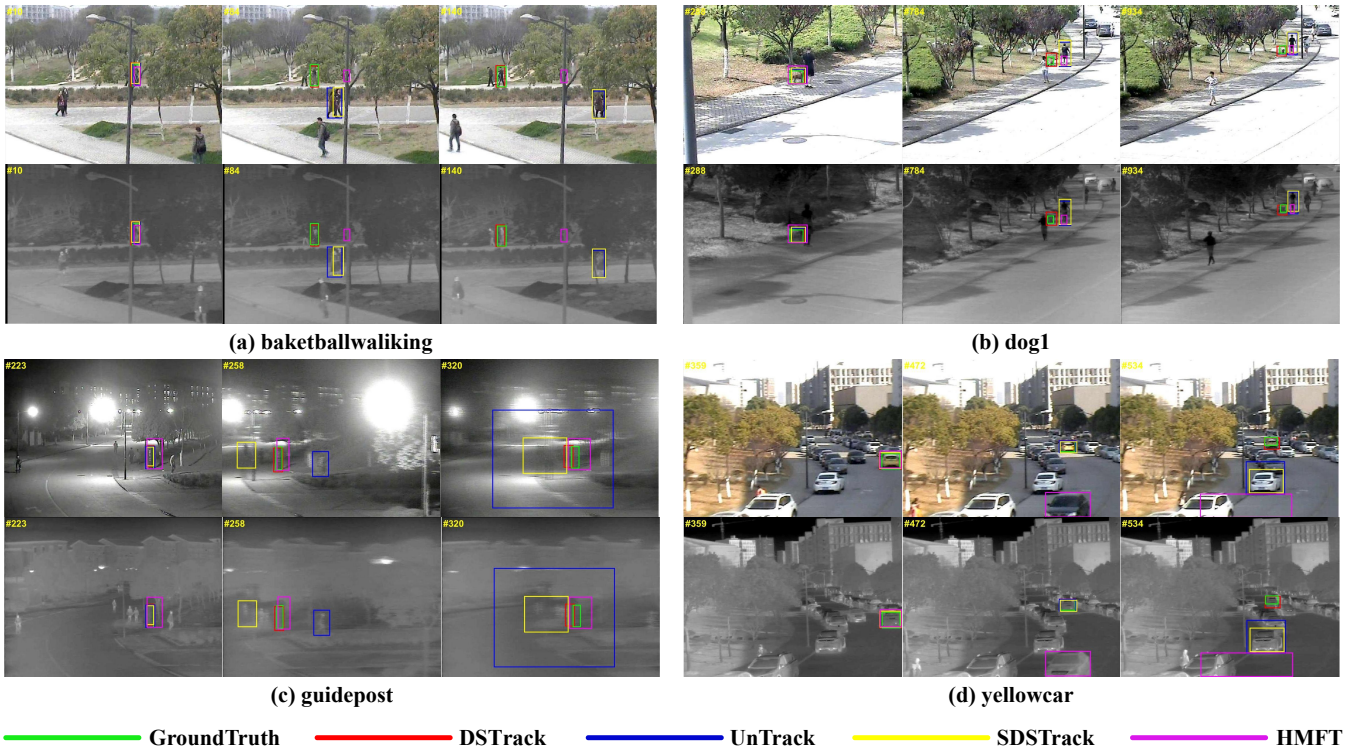

Fig. 8. Visualized comparisons of DSTrack with UnTrack [20], SDSTrack [21] and HMFT [12] on four sequences from the RGBT234 [10] dataset.

important role in improving the performance of tracking tasks.

4) Visualization of Heatmap: To further demonstrate the effectiveness of MDENet, we visualize the heatmaps before and after adding MDENet. As shown in Fig. 7, in the bikeboyright and rightbluewhite sequences, distractor objects interfere the target appearance in both scenarios, which is difficult for ToMP method to identify and understand the target, resulting in inaccurate multi-peak responses in the heatmap. In contrast, DSTrack effectively improves the discriminability of target features, accurately focusing on the target region and generating a single-peak response. In the bikefromlight sequence, strong illumination and dark night conditions severely degrade both RGB and TIR modality quality, DSTrack still achieves an accurate single-peak response. Similarly, in the whitegirl sequence, DSTrack produces a more concentrated heatmap response compared to ToMP, demonstrating superior target localization. Overall, these visualizations validate the performance improvements brought by our method in handling complex scenarios.

5) Visualization of Tracking Results: As shown in Fig. 8, we select four representative sequences from the RGBT234 dataset to display some tracking results of DSTrack and other state-of-the-art RGB-T trackers. For example, in the basketballwalking sequence, when the target is frequently occluded, other trackers fail to maintain stable tracking, while DSTrack consistently locates the target. In the dog1 long sequence, as the small target gradually moves away, DSTrack

still achieves precise tracking, whereas other methods have significant errors. In the guidepost sequence, when other trackers lose the target due to camera shake, DSTrack still tracks the target stably, indicating better robustness.

IV. Conclusion

In this paper, we have proposed DSTrack, a novel Dynamic Enhancement Network with switch for RGB-T Tracking. This framework has introduced a dynamic switch mechanism that adaptively selected optimal model configurations based on scenario characteristics, enabling structural flexibility for complex environment adaptation. Besides, we have designed a multi-domain enhancement network that synergistically combined intra-modal and inter-modal enhancement modules. The intra-modal component employed a tri-domain enhancement strategy (channel, spatial, and frequency domains) to amplify discriminative target features, while the inter-modal module established complementary cross-attention links to compensate for single-modal deficiencies. Our experimental results have demonstrated the superiority of DSTrack over state-of-the-art methods, showing its potential in handling real-world tracking tasks across several practical applications.

In the future, we plan to optimize computational efficiency by introducing more lightweight strategies in both the switch selection mechanism and multi-domain enhancement networks. Such as reducing redundant parameters through pruning, or incorporating more compact enhancement modules via distillation learning.

References

[1] H. Zhang *et al.*, "A comprehensive review of RGBT tracking," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–23, 2024.

[2] D. K. Jain, X. Zhao, C. Gan, P. K. Shukla, A. Jain, and S. Sharma, "Fusion-driven deep feature network for enhanced object detection and tracking in video surveillance systems," *Inf. Fusion*, vol. 109, p. 102429, 2024.

[3] Y. Hu, Y. Liu, G. Cao, and J. Wang, "GLFNet: An RGB-T crowd counting network based on globallocal multimodal feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–18, 2025.

[4] P. Zhang, Y. Li, Y. Zhuang, J. Kuang, X. Niu, and R. Chen, "Multi-level information fusion with motion constraints: Key to achieve high-precision gait analysis using low-cost inertial sensors," *Inf. Fusion*, vol. 89, pp. 603–618, 2023.

[5] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5613–5625, 2021.

[6] L. Lei and X. Li, "RGB-T tracking with frequency hybrid awareness," *Image Vis. Comput.*, vol. 152, p. 105330, 2024.

[7] Y. Luo, X. Guo, M. Dong, and J. Yu, "RGB-T tracking based on mixed attention," 2023, *arXiv:2304.04264*.

[8] L. Liu, C. Li, Y. Xiao, and J. Tang, "Quality-aware RGBT tracking via supervised reliability learning and weighted residual guidance," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 3129–3137.

[9] 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*, 2017, pp. 1856–1864.

[10] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, p. 106977, 2019.

[11] C. Li *et al.*, "LasHeR: A large-scale high-diversity benchmark for RGBT tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 392–404, 2022.

[12] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal UAV tracking: A large-scale benchmark and new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8876–8885.

[13] J. Mei, D. Zhou, J. Cao, R. Nie, and K. He, "Differential reinforcement and global collaboration network for RGBT tracking," *IEEE Sensors J.*, vol. 23, no. 7, pp. 7301–7311, 2023.

[14] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient RGB-T tracking via cross-modality distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5404–5413.

[15] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 9516–9526.

[16] T. Hui *et al.*, "Bridging search region interaction with template for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13 630–13 639.

[17] M. Feng and J. Su, "RGBT image fusion tracking via sparse trifurcate transformer aggregation network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2024.

[18] L. Liu, C. Li, Y. Xiao, R. Ruan, and M. Fan, "RGBT tracking via challenge-based appearance disentanglement and interaction," *IEEE Trans. Image Process.*, vol. 33, pp. 1753–1767, 2024.

[19] L. Hong *et al.*, "OneTracker: Unifying visual object tracking with foundation models and efficient tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 19 079–19 091.

[20] Z. Wu *et al.*, "Single-model and any-modality for video object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 19 156–19 166.

[21] X. Hou *et al.*, "SDSTrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 26 541–26 551.

[22] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 2252–2261.

[23] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware RGBT tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 222–237.

[24] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for RGBT tracking," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2831–2838.

[25] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, "Duality-gated mutual condition network for RGBT tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4118–4131, 2025.

[26] X. Wang *et al.*, "MFGNet: Dynamic modality-aware filter generation for RGB-T tracking," *IEEE Trans. Multimedia*, vol. 25, pp. 4335–4348, 2023.

[27] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[28] C. Mayer *et al.*, "Transforming model prediction for tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8721–8730.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[31] 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)*, 2019, pp. 91–99.

[32] M. Kristan *et al.*, "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 2206–2241.

[33] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time RGB-T tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 2714–2729, 2021.

[34] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8122–8131.