RGBT Tracking via All-layer Multimodal Interactions with Progressive Fusion Mamba

Andong Lu[1], Wanyu Wang[2], Chenglong Li[2*], Jin Tang[1], Bin Luo[1†]

[1]School of Computer Science and Technology, Anhui University
[2]School of Artificial Intelligence, Anhui University
adlu_ah@foxmail.com, 2878065167@qq.com, lcl1314@foxmail.com, tangjin@ahu.edu.cn, luobin@ahu.edu.cn

Abstract

Existing RGBT tracking methods often design various interaction models to perform cross-modal fusion of each layer, but can not execute the feature interactions among all layers, which plays a critical role in robust multimodal representation, due to large computational burden. To address this issue, this paper presents a novel All-layer multimodal Interaction Network, named AINet, which performs efficient and effective feature interactions of all modalities and layers in a progressive fusion Mamba, for robust RGBT tracking. Even though modality features in different layers are known to contain different cues, it is always challenging to build multimodal interactions in each layer due to struggling in balancing interaction capabilities and efficiency. Meanwhile, considering that the feature discrepancy between RGB and thermal modalities reflects their complementary information to some extent, we design a Difference-based Fusion Mamba (DFM) to achieve enhanced fusion of different modalities with linear complexity. When interacting with features from all layers, a huge number of token sequences (3840 tokens in this work) are involved and the computational burden is thus large. To handle this problem, we design an Order-dynamic Fusion Mamba (OFM) to execute efficient and effective feature interactions of all layers by dynamically adjusting the scan order of different layers in Mamba. Extensive experiments on four public RGBT tracking datasets show that AINet achieves leading performance against existing state-of-the-art methods. We will release the code upon acceptance of the paper.

Introduction

RGBT tracking aims to leverage the complementary information of visible light (RGB) and thermal infrared (TIR) images to predict the location and size of an object. By combining the penetration capability and nighttime sensitivity of TIR images with the rich color and texture of RGB images under favorable lighting conditions, RGBT tracking has attracted significant research attention. Numerous innovative works (Zhu et al. 2021; Hui et al. 2023; Liu et al. 2023a; Wang et al. 2024a; Cao et al. 2024; Lu et al. 2024) have been proposed to improve the robustness and accuracy of RGBT tracking.

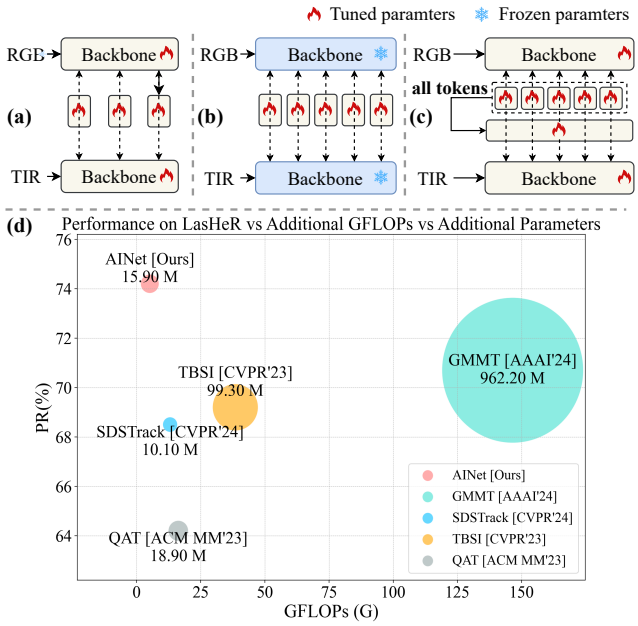

Figure 1: Comparison with existing RGBT tracking methods. (a) Interactions between specific layers, with joint fine-tuning of the entire backbone. (b) Interactions between all corresponding layers, with the pre-trained backbone being frozen. (c) Interactions between all corresponding layers, and interactions among all layers, with joint fine-tuning with the backbone. (d) Performance comparison on LasHeR, and comparison of additional parameters and GFLOPs.

As a multimodal vision task, most RGBT tracking studies focus on modality interaction modules, and these approaches can be broadly classified into two categories. One category involves building complex interaction modules (Hui et al. 2023; Fan et al. 2024; Tang et al. 2024) with powerful representation to achieve inter-modal interactions. For example, (Hui et al. 2023) employ six cross-attention blocks and extra fused template features for inter-modal feature interactions. Similarly, (Fan et al. 2024)

*Corresponding author
†Corresponding author

stack multiple Swin-Transformer (Liu et al. 2021) blocks to perform self-attention between two modality features. However, due to the significant computational burden, these methods can only interact at a limited number of specific layers, as shown in Fig. 1(a). Another category involves designing lightweight interaction modules and adopting the strategy of interacting between all corresponding layers to achieve inter-modal interactions, as illustrated in Fig. 1 (b). For instance, (Cao et al. 2024) deploy lightweight adapters in all layers to perform bi-directional inter-modal interactions. Similar interaction strategies are also demonstrated in (Hou et al. 2024; Zhu et al. 2023). However, these methods have restricted interaction capabilities and representation capacity due to their extremely low parameter count. As a result, existing methods struggle to balance interaction capability and efficiency when constructing multimodal interactions between all corresponding layers.

In addition, features from different layers show significant complementarity: low-level features provide detailed texture information, while high-level features capture abstract and semantic content. Nevertheless, existing methods adopt CNNs or Transformer networks, where the small receptive field of CNNs hinders the modeling of global information between modalities, and the Transformer is hard to interact with information from multi-layer features due to the $O(N^2)$ computational complexity. Hence, these architectural limitations prevent previous RGBT tracking methods from achieving comprehensive interaction of all layer information, which is essential for robust RGBT tracking.

To address these issues, we propose a novel All-layer multimodal Interaction Network named AINet, as shown in Fig. 1 (c), for robust RGBT tracking. In particular, modality feature interaction aims to leverage the mutual enhancement and complementarity of two modalities, while complementary information reflected in their differences. Thus, we design a difference-based fusion Mamba (DFM) with linear complexity, which not only can model modality differences to enhance each modality feature, but also can be efficiently applied to each layer. When performing feature interactions across all layers, handling a large number of token sequences (3840 tokens) results in a significant computational burden for existing networks. Although the standard Mamba network is efficient, it is prone to forgetting early token information due to the inherent properties of causal models. To mitigate this issue, we design an Order-dynamic Fusion Mamba (OFM) module, which dynamically adjusts the scan order of different layers based on the input to alleviate layer information forgetting. Extensive experiments on four public RGBT tracking benchmarks show that AINet significantly outperforms existing state-of-the-art methods in both performance and efficiency, as depicted in Fig. 1 (d). Our main contributions are summarized as follows:

- We propose a novel all-layer multimodal interaction network for RGBT tracking. It conducts multimodal interaction of each layer and all layer interaction in a progressive fusion Mamba. To the best of our knowledge, we are the first to introduce the Mamba network in RGBT tracking.
- We design a difference-based fusion Mamba, which achieves inter-modal enhanced fusion by modelling inter-modal differences to capture complementary information, and efficiently applies it to each layer.
- We design an order-dynamic fusion Mamba, which implements all-layer feature interaction with an input-aware dynamic scanning scheme to mitigate information forgetting of early input tokens.
- Extensive experiments on four RGBT benchmarks demonstrate that AINet achieves new state-of-the-art results while maintaining a controllable number of parameters and computational load.

Related Work

RGBT Tracking

In recent years, the field of RGBT tracking has made significant progress, with many impressive works emerging. Existing methods can be roughly divided into two categories based on their feature fusion strategies. One category involves late fusion, which takes place after the main feature extraction backbone. For instance, (Zhang et al. 2019a) embed the multimodal feature concatenation process into the framework of a strong tracker DiMP (Bhat et al. 2019) for RGBT tracking. (Peng et al. 2023) utilize two-stream convolutional network with increasing coupling filters to extract both common and individual features, and ultimately achieves fusion through a simple channel concatenation. The second category involves modality interaction during the feature extraction phase. For example, (Hui et al. 2023) build a template-bridged cross-attention module between the RGB and TIR search areas to promote thorough modality fusion. (Cao et al. 2024) propose a bi-directional adapter to perceive the dominant modality changes in dynamic scenes and adaptively fuse multimodal information. However, existing methods are limited by computational costs and cannot utilize information from all layers. Our method can fully leverage all-layer multimodal information while keeping a small number of parameters and low computational resource consumption.

Vision State Space Model

State space models (SSMs) (Gu et al. 2021; Nguyen et al. 2022; Zhang et al. 2023a) derived from classical control theory connects the input and output sequences through hidden states. Recently, Mamba (Gu and Dao 2023) has been widely applied in various fields due to its selective mechanism and efficient hardware acceleration design. (Zhu et al. 2024) expand to visual tasks for the first time by bidirectional sequence modeling. (Guo et al. 2024) propose a residual state space block, using convolution and channel attention to enhance the performance of Vanilla Mamba, and achieved success in the field of image restoration. (Zhang et al. 2024) integrate a selective scanning mechanism into the motion generation task, constructing HTM and BSM modules to handle temporal motion data and bidirectionally capture the channel-wise flow of hidden information within the latent pose. However, the application of Mamba in RGBT tracking has not yet been sufficiently explored. In this work, we leverage Mamba for modality enhancement and all-layer fusion, exploring the potential of Mamba in RGBT tracking.

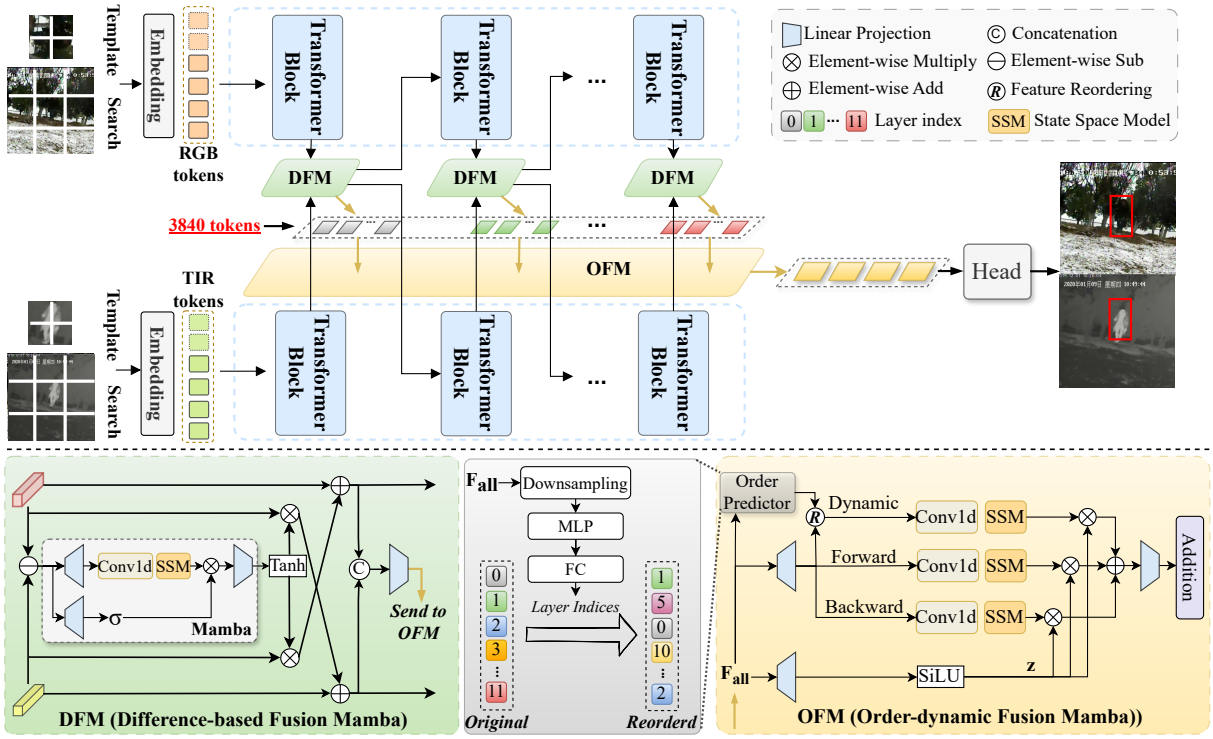

Figure 2: The overall architecture of our proposed AINet. Firstly, RGB and TIR images are embedded as tokens and fed into Transformer blocks for joint feature extraction and relationship modeling between search and template images. Following each block, the tokens from both modalities are processed by the DFM for difference information enhancement and then returned to the backbone. Meanwhile, the fusion features at each layer are cascaded and fed into the OFM for all-layer interaction. Finally, the output features from the OFM are sent to the tracking head for target localization.

Methodology

In this paper, we propose a novel All-layer multimodal Interaction Network, named AINet, for RGBT tracking, which performs efficient and effective feature interactions of all modalities and layers in a progressive fusion Mamba. In particular, AINet achieves multimodal interactions at each layer by the designing difference-based fusion Mamba, and it employs an order-dynamic fusion Mamba to establish all-layer interactions. Next, we first review the mamba, then introduce the overall architecture of AINet, and finally, we describe in detail the two fusion Mamba architectures.

To visually demonstrate the necessity of applying all layer features, we perform a visual analysis of the final fusion features that incorporate different numbers of layer features, as shown in Fig. 3. It can be observed that as the number of layers increases, the model's response to the target becomes more comprehensive and focused, validating the necessity and importance of applying all layer features.

Preliminaries

The state-space sequence model (SSM) (Gu et al. 2021) and Mamba (Gu and Dao 2023) are inspired by continuous linear systems, where a one-dimensional function or sequence, denoted as $x(t) \in \mathbb{R}$, is mapped to $x(t) \in \mathbb{R}$ through a hidden states $h(t) \in \mathbb{R}^N$. The models can be represented by

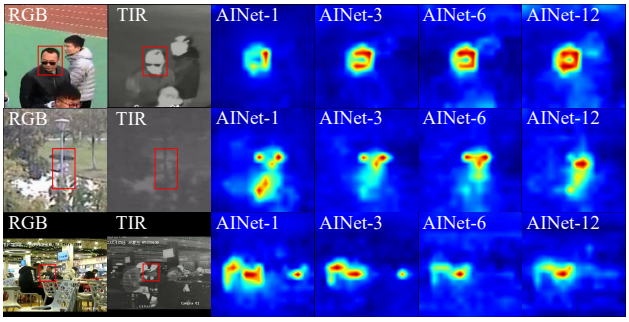

Figure 3: Illustration of fusion feature visualization with different layers applied. Here, "n" in AINet-"n" represents the number of layers applied.

linear ordinary differential equations (ODEs) as follows:

$$h^{'}(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t),\ y(t) = \boldsymbol{C}h(t), \tag{1}$$

where N is the state size, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$. Specifically, Δ denotes the timescale parameter to transform the continuous parameters A, B to discrete parameters A, B. The commonly used method for discretization is the zero-order hold (ZOH) rule, defined as follows:

$$\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}),\ \overline{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}. \tag{2}$$

Then, the discretized version of Eq. 1 with step size Δ can be rewritten as:

$$h_t = \overline{\boldsymbol{A}} h_{t-1} + \overline{\boldsymbol{B}} x_t,\ y_t = \boldsymbol{C} h_t. \tag{3}$$

Finally, the models compute output $\mathbf{y}$ through a global convolution within a structured convolutional kernel $\overline{\boldsymbol{K}}$.

$$\overline{\boldsymbol{K}} = (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}}, \ldots, \boldsymbol{C}\overline{\boldsymbol{A}}^{\boldsymbol{M-1}}\overline{\boldsymbol{B}}),\ \mathbf{y} = \mathbf{x} * \overline{\boldsymbol{K}}, \tag{4}$$

where $\boldsymbol{M}$ denotes the sequence length of $\mathbf{x}$. In contrast to traditional SSMs, Mamba introduces the Selective Scanning Mechanism (S6), which allows selective focus on what is important in long contexts. In this work, we extend the Mamba module to support all-layer multimodal interactions.

Overall Architecture

Inspired by the success of ViT (Dosovitskiy et al. 2021) in tracking tasks, we follow the OSTrack (Ye et al. 2022) framework and extend its backbone to adapt to multimodal input. As shown in Fig. 2, our AINet incorporates a two-stream encoder structure, sharing the same parameters for both the RGB and TIR modalities. Additionally, it includes a set of difference-based fusion Mamba (DFM), an order-dynamic fusion Mamba (OFM), and a prediction head. Specifically, AINet first processes the search and template frames of the given RGB and TIR modalities through the patch and position embedding layers, obtaining the initial token pairs for each modality. The search and template tokens from each modality are then concatenated along the token dimension to form RGB tokens x_0^{rgb} and TIR tokens x_0^{tir}, which are fed into the ViT blocks for feature extraction and joint relationship modeling. Since the DFM is embedded in each layer to facilitate modality interactions, for the i-th layer block, it learns to integrate enhanced modality features from the output of the previous DFM, as described below.

$$(\hat{x}_i^{rgb}, \hat{x}_i^{tir}, x_i^{dfm}) = \mathcal{F}_i^{DFM}(x_i^{rgb}, x_i^{tir}), i \in [1, N], \tag{5}$$

where N refers to the total number of blocks. $\hat{x}_i^{rgb}$ and $\hat{x}_i^{tir}$ represent the enhanced RGB and TIR modality features, respectively. $\mathcal{F}_i^{DFM}$ and x_i^{dfm} denote the DFM and the fused features, respectively, at the i-th layer. When all blocks are executed, x_i^{dfm} from all layers are concatenated along the token dimension and fed into the OFM, represented as $\mathcal{F}_i^{OFM}$, for all layers to interact. Then, the prediction head $\mathcal{H}$ predicts the tracking result $\mathcal{B}$ based on the output of the OFM.

$$\mathcal{B} = \mathcal{H}(\mathcal{F}^{OFM}[x_1^{dfm}, x_2^{dfm}, \ldots, x_N^F]), \tag{6}$$

where $[\cdot]$ refers to concatenation along the token dimension.

Difference-based Fusion Mamba (DFM)

Visible and thermal infrared modalities capture complementary object properties due to different imaging principles, which implies that the differences between modalities often contain complementary information. Nevertheless, current advanced approaches mainly adopt Transformer networks with long-range modeling capabilities to directly interact the features of the modalities, which ignores the explicit utilization of modality differences and limits multimodal interactions at each layer due to high secondary computational overhead. To this end, we design a difference-based fusion Mamba (DFM) with linear complexity, as depicted in Fig. 2 (b), which can be employed to model modality differences at each layer to enhance modal representation. To capture inter-modal differences and enrich feature learning, DFM employs the principles of differential amplifier circuits, which suppresses common-mode signals and amplifies differential ones.

In particular, we obtain the modality difference feature x_i^d by subtracting between modal features in the same layer. Since the difference feature contains both useful information and noise, x_i^d is fed to the Mamba to suppress noise while enhancing useful information. Next, x_i^d, processed by the activation function, is multiplied element-wise with x_i^{rgb} and x_i^{tir} to obtain the difference compensation features for each modality. Finally, these compensated features are added back to x_i^{rgb} and x_i^{tir} to obtain the enhanced features $\hat{x}_i^{rgb}$ and $\hat{x}_i^{tir}$. The process is summarized as follows:

$$\begin{aligned} \hat{x}_i^{rgb} &= x_i^{rgb} + x_i^{tir} \times \tau(Mamba(x_i^d)), \\ \hat{x}_i^{tir} &= x_i^{tir} + x_i^{rgb} \times \tau(Mamba(x_i^d)), \end{aligned} \tag{7}$$

where τ denotes the $tanh$ function. Subsequently, the enhanced modality features are processed to obtain the fused feature of the current layer:

$$x_i^{dfm} = \tau(LN([x_i^{rgb}, x_i^{tir}] \cdot W_i)), \tag{8}$$

where $W_i \in \mathbb{R}^{2C \times C}$ is a linear layer with reduced channel dimension, and LN represents layer normalization.

Order-dynamic Fusion Mamba (OFM)

The strong complementarity between different feature layers in deep networks has been proven in many visual tasks (Lin et al. 2017; Pang et al. 2020; Liu et al. 2023b; Wang et al. 2024b). However, no current method applies all feature layers to RGBT tracking, primarily because existing methods use Transformers for feature interactions.

To this end, we design an Order-dynamic Fusion Mamba (OFM) to efficiently and effectively interact with features from all layers by dynamically adjusting the scan order of different layers in Mamba.

In particular, We first concatenate the output features x_i^{dfm} of each DFM layer along the token dimension to form a long token sequence F_{all} containing features from all layers. We then input F_{all} to the OFM and perform the following forward and backward scanning modeling process:

$$\begin{aligned} F_{all}^{forward} &= \mathcal{SSM}([x_1^{dfm}, x_2^{dfm}, \ldots, x_N^{dfm}], W_c), \\ F_{all}^{backward} &= \mathcal{SSM}([x_N^{dfm}, x_{N-1}^{dfm}, \ldots, x_1^{dfm}], W_c), \end{aligned} \tag{9}$$

where W_c represents a 1D convolution layer and $\mathcal{SSM}$ denotes the selective scanning model. However, only performing forward and backward scans can potentially ignore the first and last layer tokens. Therefore, we propose an order-dynamic scanning scheme that allows the scanning process

to start and end at any layer. This innovative design enables OFM to rank the importance of different layer features based on the input data.

In the order-dynamic scanning modeling, F_{all} is first downsampled ($\mathcal{D}$) to the specified dimension and then fed into a multi-layer perceptron ($\mathcal{MLP}$) and a fully connected layer ($\mathcal{FC}$) to predict an index covering the scanning order of all layers. The process is expressed as follows:

$$index = \mathcal{FC}(\mathcal{MLP}(\mathcal{D}(F))). \tag{10}$$

Then, the OFM reorders the long token sequence by the $index$. Thus, the dynamic ordering scan modeling can be formulated as follows:

$$\begin{aligned} x_{all}^{order} &= \{[x_1^{dfm}, x_2^{dfm}, \ldots, x_N^{dfm}], index\}, \\ F_{all}^{dynamic} &= \mathcal{SSM}(x_{all}^{order}, W_c), \end{aligned} \tag{11}$$

where $\{\cdot, index\}$ represents the input sequence ordered according to the given index, and x_{all}^{order} denotes the ordered result. Next, a simple gating strategy in Mamba fuses the results of three-scan modeling. Finally, all layer features are aggregated by element-wise addition and fed into the tracking head.

Experiment

Implementation Details

We implement our AINet based on the PyTorch and train it on single NVIDIA RTX 4090 GPU. We follow the hyper-parameter settings of the baseline model (Ye et al. 2022) for the loss function. For parameter initialization, we utilize the pretrained model provided by DropTrack (Wu et al. 2023). For each sequence in the training set, we collect training samples and apply standard data augmentation operations, including rotation, translation, and gray-scaling, following the data processing scheme of the base tracker (Ye et al. 2022). During training, we use the AdamW (Loshchilov and Hutter 2017) optimizer with a weight decay of 10^{-4}, and set the batch size and learning rate to 16 and 10^{-4}, respectively. The entire network is trained end-to-end over 15 epochs, with each epoch providing 6×10^4 pairs of samples. We use the LasHeR training set to train our network, which is used to evaluate RGBT210 (Li et al. 2017), RGB234 (Li et al. 2019) and LasHeR (Li et al. 2021). For the evaluation of VTUAV (Zhang et al. 2022), we utilize the training set from VTUAV as the training data.

Quantitative Comparison

We evaluate our proposed AINet on four popular RGBT tracking benchmarks: RGBT210, RGBT234, LasHeR and VTUAV, and compare the performance with 20 state-of-the-art RGBT trackers. We adopt the Precision Rate (PR), Success Rate (SR), and Normalized Precision Rate (NPR) from One-Pass Evaluation (OPE) as metrics for quantitative performance measurement, which are commonly used in current RGBT tracking tasks. The effectiveness of our proposed AINet is demonstrated in Table 1, which summarizes the comparison results.

Evaluation on RGBT210 dataset. RGBT210 is a challenging dataset that contains 210 pairs of RGBT video sequences, totaling approximately 210K frames, and provides annotations for 12 different challenge attributes. As shown in Table 1, AINet achieves the best performance, outperforming the current state-of-the-art QAT and TATrack by 0.7%/2.9% and 2.2%/3.0% in PR/SR, respectively.

Evaluation on RGBT234 dataset. RGBT234 is one of the most influential and extensively evaluated RGBT tracking dataset, containing 234 pairs of aligned RGBT videos, amounting to approximately 233.4K frames. As shown in Table 1, we evaluate AINet against 20 state-of-the-art trackers on RGBT234 dataset. Compared with QAT and GMMT, the second best-performing algorithms in PR and SR respectively, our method shows a performance advantage of 0.8%/2.9% in PR/SR and 1.3%/2.6% in PR/SR respectively.

Evaluation on LasHeR dataset. LasHeR is the largest and most challenging RGBT tracking dataset, containing 1,224 aligned RGBT video sequences totaling approximately 734.8K frames. It also provides annotations for 19 challenge attributes, such as HI (High Illumination) and AIV (Abrupt Illumination Variation), significantly raising the dataset's challenge.

1) Overall Comparison. As shown in Table 1, we compare our method with 17 state-of-the-art trackers using PR, NPR, and SR metrics on the LasHeR dataset. The results demonstrate that our method significantly outperforms other trackers. Specifically, compared to the most powerful tracker, GMMT, our method achieves improvements of 3.5%/3.1%/2.5% in PR/NPR/SR. Compared to the unified frameworks OneTracker, Un-Track, and SDSTrack, we achieve performance improvements of 7.0%/5.3%, 7.5%/5.5%, and 7.7%/6.0% in PR/SR, respectively. These results fully demonstrate the superiority of our method.

2) Challenge-based Comparison. We also present the results of AINet against the most advanced RGBT trackers, including SDSTrack (Hou et al. 2024), GMMT (Tang et al. 2024), QAT (Liu et al. 2023a), and TBSI (Hui et al. 2023), on different challenge subsets. The evaluation results are shown in Fig. 4, where each corner represents the attributes of the challenge subset and the highest and lowest performance under that attribute. The challenge subsets include no occlusion (NO), partial occlusion (PO), total occlusion (TO), hyaline occlusion (HO), motion blur (MB), low illumination (LI), high illumination (HI), abrupt illumination variation (AIV), low resolution (LR), deformation (DEF), background clutter (BC), similar appearance (SA), camera moving (CM), thermal crossover (TC), frame lost (FL), out-of-view (OV), fast motion (FM), scale variation (SV), and aspect ratio change (ARC). The results show that AINet achieves the best performance in almost all subsets, especially with significant improvements in scenarios like MB, DEF, and FL, proving that AINet has great potential in various complex tracking scenarios.

Evaluation on VTUAV dataset. VTUAV collects RGBT data from UAV scenarios, expanding the application of RGBT tracking. It contains 500 aligned RGBT video sequences with up to 1.7 million frames. We focus our experiments on its short-term tracking subset. As shown in Table 1,

Methods	Pub. Info.	Backbone	RGBT210		RGBT234		LasHeR			VTUAV		FPS
			PR↑	SR↑	PR↑	SR↑	PR↑	NPR↑	SR↑	PR↑	SR↑	↑
mfDiMP (Zhang et al. 2019b)	ICCVW 2019	ResNet−50	78.6	55.5	–	–	44.7	39.5	34.3	67.3	55.4	10.3
CAT (Li et al. 2020)	ECCV 2020	VGG−M	79.2	53.3	80.4	56.1	45.0	39.5	31.4	–	–	20
ADRNet (Zhang et al. 2021)	IJCV 2021	VGG-M	–	–	80.7	57.0	–	–	–	62.2	46.6	25
APFNet (Xiao et al. 2022)	AAAI 2022	VGG−M	–	–	82.7	57.9	50.0	43.9	36.2	–	–	1.3
DMCNet (Lu et al. 2022)	TNNLS 2022	VGG−M	79.7	55.5	83.9	59.3	49.0	43.1	35.5	–	–	2.3
ProTrack (Yang et al. 2022)	ACM MM 2022	ViT−B	–	–	78.6	58.7	50.9	–	42.1	–	–	30
HMFT (Zhang et al. 2022)	CVPR 2022	ResNet−50	78.6	53.5	78.8	56.8	–	–	–	75.8	62.7	30.2
MFG (Wang et al. 2022)	TMM 2022	ResNet−18	74.9	46.7	75.8	51.5	–	–	–	–	–	–
DFNet (Peng, Zhao, and Hu 2022)	TITS 2022	VGG−M	–	–	77.2	51.3	–	–	–	–	–	–
DRGCNet (Mei et al. 2023)	IEEE SENS J 2023	VGG−M	–	–	82.5	58.1	48.3	42.3	33.8	–	–	4.9
CMD (Zhang et al. 2023b)	CVPR 2023	ResNet−50	–	–	82.4	58.4	59.0	54.6	46.4	–	–	30
ViPT (Zhu et al. 2023)	CVPR 2023	ViT−B	–	–	83.5	61.7	65.1	–	52.5	–	–	–
TBSI (Hui et al. 2023)	CVPR 2023	ViT−B	85.3	62.5	87.1	63.7	69.2	65.7	55.6	–	–	36.2
QAT (Liu et al. 2023a)	ACM MM 2023	ResNet−50	86.8	61.9	88.4	64.4	64.2	59.6	50.1	80.1	66.7	22
TATrack (Wang et al. 2024a)	AAAI 2024	ViT−B	85.3	61.8	87.2	64.4	70.2	66.7	56.1	–	–	26.1
BAT (Cao et al. 2024)	AAAI 2024	ViT−B	–	–	86.8	64.1	70.2	–	56.3	–	–	–
GMMT (Tang et al. 2024)	AAAI 2024	ViT−B	–	–	87.9	64.7	70.7	67.0	56.6	–	–	–
OneTracker (Hong et al. 2024)	CVPR 2024	ViT−B	–	–	85.7	64.2	67.2	–	53.8	–	–	–
Un-Track (Wu et al. 2024)	CVPR 2024	ViT−B	–	–	84.2	62.5	66.7	–	53.6	–	–	–
SDSTrack (Hou et al. 2024)	CVPR 2024	ViT−B	–	–	84.8	62.5	66.5	–	53.1	–	–	20.9
AINet (256×256)	–	ViT−B	86.8	64.1	89.1	66.8	73.0	69.0	58.2	87.1	74.5	**38.1**
AINet (384×384)	–	ViT−B	**87.5**	**64.8**	**89.2**	**67.3**	**74.2**	**70.1**	**59.1**	**88.0**	**75.3**	37.5

Table 1: The PR, NPR, and SR scores (%) of various trackers on five datasets. **Bold** indicates the best result.

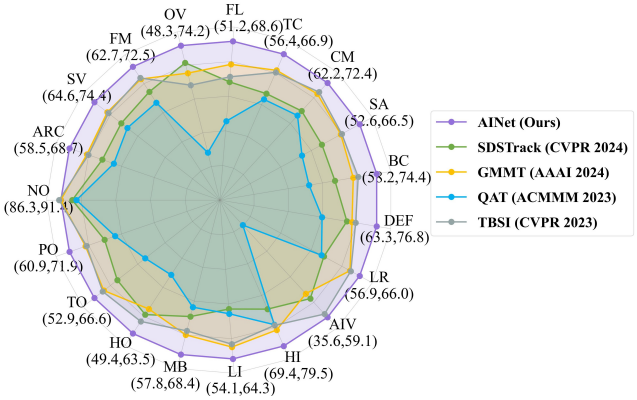

Figure 4: Precision Rate (PR) of challenge attributes on LasHeR. The axes of each attribute have been normalized.

AINet achieves a PR/SR performance of 88.0%/75.3%, outperforming four other trackers. Compared to QAT (Liu et al. 2023a), the second-best algorithm on this dataset, AINet shows an advantage of 7.9%/8.6% in PR/SR. These results shows the advantages of AINet under the UAV perspective.

Ablation Study

Component analysis. In Table 2, we conduct several ablation studies on the LasHeR dataset to verify the effectiveness of key components in our AINet.

Baseline denotes the removal of DFM and OFM modules from our method, while maintaining consistent training data and losses.

w/ DFM indicates that each layer in the Baseline backbone network is equipped with a DFM module, achieving improvements of 1%/0.8%/0.5% in PR/NPR/SR, respectively. This experiment shows that difference-based fusion Mamba is effective.

Methods	Resolution	PR↑	NPR↑	SR↑
Baseline	256×256	71.1	67.5	56.9
w/ DFM	256×256	72.1	68.3	57.4
w/ OFM	256×256	72.2	68.0	57.3
w/ DFM OFM*	256×256	72.5	68.6	57.8
w/ DFM OFM	256×256	73.0	69.1	58.2
w/ DFM OFM	384×384	**74.2**	**70.1**	**59.1**

Table 2: Quantitative comparison of different variants of our method on the LasHeR dataset. OFM* indicates the removal of the order-dynamic scanning scheme within the OFM.

w/ OFM denotes that each layer in the backbone uses simple feature addition to obtain fused features and is equipped with our proposed OFM module, achieving improvements of 1.1%/0.5%/0.4% in PR/NPR/SR, respectively. This experiment shows that order-dynamic fusion Mamba is effective.

w/ DFM OFM represents applying both our proposed DFM and OFM modules in the Baseline, achieving a clear performance improvement of 1.9%/1.6%/1.3% in PR/NPR/SR metrics, respectively. This experiment demonstrates the effectiveness of our proposed progressive fusion Mamba. In addition, we remove our designed order-dynamic scanning scheme in ***w/ DFM OFM*** and denote it as ***w/ DFM OFM****. The results show a certain decrease, proving the effectiveness and necessity of order-dynamic scanning.

Impact of different resolutions. Increasing the resolution of input images for performance improvement has

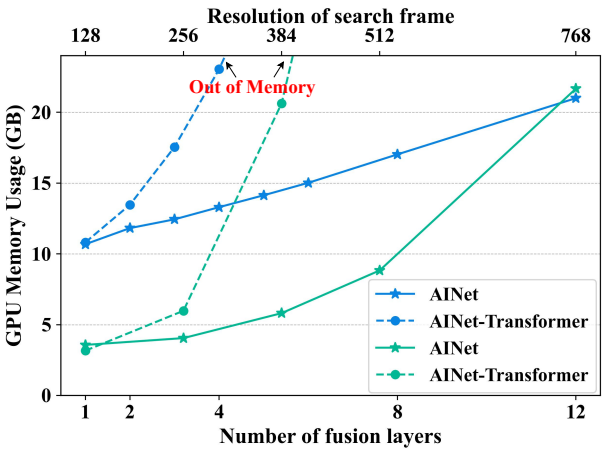

Figure 5: Comparison of GPU memory usage between our method and a transformer-based approach under variations in layer count and resolution. **Blue** line and **Green** line respectively indicate changes in resolution and layer count.

Variants	Layer index	PR↑	NPR↑	SR↑
AINet-1	11	71.4	67.4	56.9
AINet-3	0,6,11	71.9	68.1	57.4
AINet-6	0,2,4,6,8,11	72.1	68.4	57.8
AINet-12	all	**73.0**	**69.1**	**58.2**

Table 3: Ablation study of applying different number of layers on the LasHeR dataset.

been proven effective in many unimodal vision tasks (Ren et al. 2015; Wang, Bochkovskiy, and Liao 2023; Cheng et al. 2024). However, existing RGBT tracking algorithms are constrained by the high computational complexity of the interaction module, preventing them from utilizing this strategy. Benefiting from the computational efficiency of Mamba, AINet introduce a larger input resolution of 384×384 to further enhance performance. As shown in Table 2, AINet achieves new state-of-the-art performance with this resolution increase, demonstrating its effectiveness across multiple datasets. In addition, we compare the GPU memory usage of AINet and its Transformer variant (AINet-Transformer) with increased resolution Fig. 5. In particular, we use cross-attention instead of DFM and self-attention instead of OFM. AINet-Transformer's memory usage grows quadratically with resolution, becoming impractical, while AINet scales linearly, maintaining efficiency. Note that the batch size is 1. These experiments fully demonstrate the efficiency advantage of AINet's interaction module.

Impact of different layers. To verify the effectiveness of utilizing all layer features for RGBT tracking, we explore the impact of employing different numbers of layers. In Table 3, we present three variants denoted as AINet-"n", where "n" represents the number of layers applied. The results show that overall performance improves as the number of layers increases. Compared to using only the last layer for fusion, involving all layers leads to a performance increase of 1.6%/1.7%/1.3% in PR/NPR/SR on LasHeR. Furthermore, we analyze resource constraints of existing interaction strategies with varying layer features. As shown in Fig. 5, the GPU memory usage of AINet-Transformer surges with additional layers, resulting in OOM (Out of Memory) at five layers. In contrast, AINet shows a linear correlation between computational resource requirements and the number of layers, allowing AINet to fully exploit information from all layers while balancing performance and efficiency.

Methods	Pub. Info.	PR↑	SR↑	Params↓	FLOPs↓	FPS↑
TBSI	CVPR 2023	69.2	55.6	99.3M	38.5G	35.7
GMMT	AAAI 2024	70.7	56.6	962.2M	146.5G	22.4
SDSTrack	CVPR 2024	66.5	53.1	**10.1M**	13.1G	18.4
Ours	–	**73.0**	**58.2**	15.9M	**5.2G**	**38.1**

Table 4: Comparison of performance, additional parameters, FLOPs, and FPS with advanced trackers on LasHeR. All algorithms' FPS are evaluated on a single RTX 4090 GPU.

Efficiency Analysis

To verify the efficiency of our method, we perform a quantitative comparison with existing state-of-the-art methods. As shown in Table 4, we present the number of parameters, computational burden compared to their respective baselines, and the inference speed of each model. Compared to the fully finetuned high-performance algorithms TBSI and GMMT, our approach is superior in all metrics.

When compared to the partially finetuned advanced method SDSTrack, despite a slight difference in the number of parameters, our approach shows a significant advantage in performance and other efficiency metrics. Notably, while our approach excels in FLOPs metrics, it underperforms in inference speed. This is attributed to the fact that Mamba is not yet adapted to the available acceleration hardware.

Conclusion

In this paper, we explore for the first time the potential of Mamba in RGBT tracking by designing a novel All-layer Interactive Network (AINet), which effectively integrates information from all layers to achieve robust tracking. The core idea of AINet lies in the progressive fusion Mamba to facilitate efficient and effective all-layer modality interactions. Specifically, we design a difference-based fusion Mamba, which enhances modality interactions at each layer by explicitly modeling the differences between modalities. Additionally, an order-dynamic fusion Mamba is designed to perform interactions across all-layer features, mitigating the risk of early information loss. Extensive experiments demonstrate the superior performance of our method in four RGBT tracking benchmarks. Currently, AINet uses ViT as the backbone for modal feature extraction, achieving strong performance but with room for efficiency improvements. To enhance efficiency, we plan to adopt Mamba as the backbone. Given the lack of pre-trained Mamba-based tracking models, we will develop a model distillation framework to build a pure Mamba-based multimodal tracking network.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62406002 and No. 62376004); and the China Postdoctoral Science Foundation (2024M760011); and the Anhui Province Science Fundation for Distinguished Young Scholars (No. 2208085J18).

References

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE international conference on computer vision*, 6182–6191.

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 927–935.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 16901–16911.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fan, H.; Yu, Z.; Wang, Q.; Fan, B.; and Tang, Y. 2024. QueryTrack: Joint-modality Query Fusion Network for RGBT Tracking. *IEEE Transactions on Image Processing*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. In *Proceedings of the IEEE European Conference on Computer Vision*.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 19079–19091.

Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. 2024. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 26551–26561.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging Search Region Interaction With Template for RGB-T Tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 13630–13639.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware RGBT tracking. In *Proceedings of the IEEE European Conference on Computer Vision*, 222–237.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of ACM International Conference on Multimedia*, 1856–1864.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, L.; Li, C.; Xiao, Y.; and Tang, J. 2023a. Quality-Aware RGBT Tracking via Supervised Reliability Learning and Weighted Residual Guidance. In *Proceddings of ACM International Conference on Multimedia*, 3129–3137.

Liu, Y.; Zhang, S.; Chen, J.; Yu, Z.; Chen, K.; and Lin, D. 2023b. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE international conference on computer vision*, 5361–5372.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, A.; Qian, C.; Li, C.; Tang, J.; and Wang, L. 2022. Duality-gated mutual condition network for RGBT tracking. *IEEE Transactions on Neural Networks and Learning Systems*.

Lu, A.; Zhao, J.; Li, C.; Xiao, Y.; and Luo, B. 2024. Breaking Modality Gap in RGBT Tracking: Coupled Knowledge Distillation. In *Proceddings of ACM International Conference on Multimedia*.

Mei, J.; Zhou, D.; Cao, J.; Nie, R.; and He, K. 2023. Differential reinforcement and global collaboration network for rgbt tracking. *IEEE Sensors Journal*, 23(7): 7301–7311.

Nguyen, E.; Goel, K.; Gu, A.; Downs, G.; Shah, P.; Dao, T.; Baccus, S.; and Ré, C. 2022. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35: 2846–2861.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9413–9422.

Peng, J.; Zhao, H.; and Hu, Z. 2022. Dynamic fusion network for RGBT tracking. *IEEE Transactions on Intelligent Transportation Systems*, 24(4): 3822–3832.

Peng, J.; Zhao, H.; Hu, Z.; Zhuang, Y.; and Wang, B. 2023. Siamese infrared and visible light fusion network for RGB-T tracking. *International Journal of Machine Learning and Cybernetics*, 14(9): 3281–3293.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28.

Tang, Z.; Xu, T.; Wu, X.; Zhu, X.-F.; and Kittler, J. 2024. Generative-based fusion mechanism for multi-modal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5189–5197.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7464–7475.

Wang, H.; Liu, X.; Li, Y.; Sun, M.; Yuan, D.; and Liu, J. 2024a. Temporal adaptive rgbt tracking with modality prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5436–5444.

Wang, K.; Tu, Z.; Li, C.; Zhang, C.; and Luo, B. 2024b. Learning Adaptive Fusion Bank for Multi-modal Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, X.; Shu, X.; Zhang, S.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2022. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking. *IEEE Transactions on Multimedia*, 25: 4335–4348.

Wu, Q.; Yang, T.; Liu, Z.; Wu, B.; Shan, Y.; and Chan, A. B. 2023. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 14561–14571.

Wu, Z.; Zheng, J.; Ren, X.; Vasluianu, F.-A.; Ma, C.; Paudel, D. P.; Van Gool, L.; and Timofte, R. 2024. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 19156–19166.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2831–2838.

Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Prompting for multi-modal tracking. In *Proceddings of ACM International Conference on Multimedia*, 3492–3500.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proceedings of the IEEE European Conference on Computer Vision*, 341–357.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; Van De Weijer, J.; and Shahbaz Khan, F. 2019a. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE international conference on computer vision workshops*, 0–0.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; and Shahbaz Khan, F. 2019b. Multi-Modal Fusion for End-to-End RGB-T Tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Zhang, M.; Saab, K. K.; Poli, M.; Dao, T.; Goel, K.; and Re, C. 2023a. Effectively Modeling Time Series with Simple Discrete State Spaces. In *The Eleventh International Conference on Learning Representations*.

Zhang, P.; Wang, D.; Lu, H.; and Yang, X. 2021. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *International Journal of Computer Vision*, 129: 2714–2729.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8886–8895.

Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; and Han, J. 2023b. Efficient RGB-T Tracking via Cross-Modality Distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5404–5413.

Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2024. Motion Mamba: Efficient and Long Sequence Motion Generation. In *Proceedings of the IEEE European Conference on Computer Vision*.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9516–9526.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Proceedings of the 41st International Conference on Machine Learning*, 62429–62442.

Zhu, Y.; Li, C.; Tang, J.; Luo, B.; and Wang, L. 2021. RGBT tracking by trident fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2): 579–592.