Exploring Historical Information for RGBE Visual Tracking with Mamba

Supplementary Material

1. Robustness Performance on FELT

Attributes. We also perform analysis of various challenging attributes, such as illumination variation, motion blur, out-of-view, etc. As shown in Figure 1, our MamTrack also achieves the best tracking performance in the most extreme scenarios, demonstrating improved robustness.

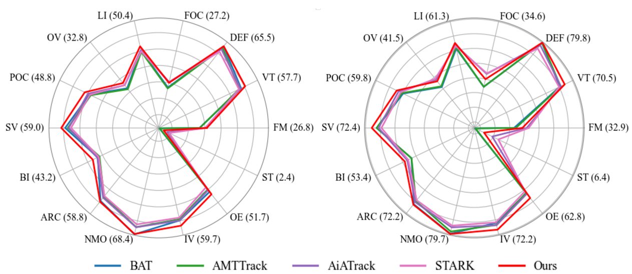


Figure 1. SR (left) and PR (right) scores of different attributes on the FELT dataset.

Visualization. Additionally, we provide a visual comparison across three representative challenging conditions on the FELT dataset to demonstrate MamTrack's ability to accurately locate the target in long-term sequences (over 3000 frames). As shown in Figure 2, our method effectively utilizes multimodal information to minimize environmental interference, such as illumination changes and background distractions. Furthermore, when the target moves out of view, historical cues enable our method to recall the target's appearance and past motion trends, thereby accurately relocating the target.

2. Details of Fusion Mamba module

	Network	SR (%)	PR(%)
A	FM Bi-Scan → FM Forward-Scan	60.5	78.0
B	FM CMS → CMS Variant	60.4	77.6
C	FM → Cross-Attention	60.8	78.1
D	FM → Bi-Adapter	61.1	78.6
E	MamTrack(Ours)	**61.6**	**79.2**

Table 1. Exploring the effectiveness of our method in the scanning scheme (A) and fusion architectures (B-D).

2.1. Impact of Scanning Scheme

The vanilla scanning scheme [3] of Mamba was originally designed for handling 1-D sequence tasks, scanning tokens in a single direction (forward), which is not suitable for 2-D vision tasks. To balance efficiency and performance, we adopt the 2-D bidirectional scanning method proposed in Vim [7] for better spatial information modeling, as shown in Figure 3. To verify the effectiveness of our bidirectional scanning scheme, we conducted additional experiments comparing the vanilla scanning scheme with ours. As shown in Table 1 A, when the model scans tokens in a single direction, the earlier tokens in the sequence fail to capture the relationships with later tokens, leading to a loss of spatial context information. Consequently, the Success Rate and Precision Rate drop by 1.1% and 1.2%, respectively.

2.2. Impact of Fusion Architecture

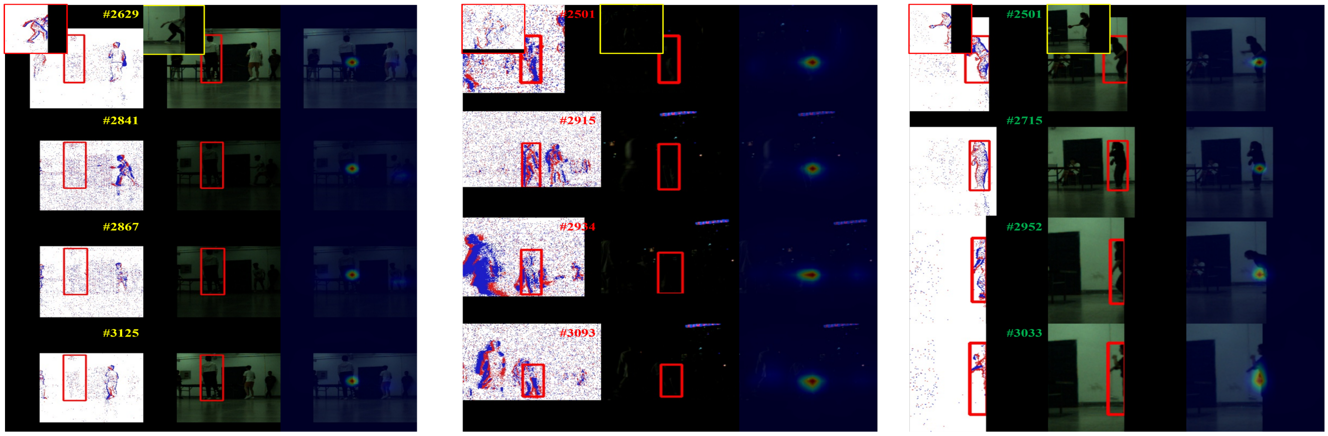
Similar to other Transformer-based multimodal approaches, several recent works [4–6] have adopted the global-aware Mamba architecture to fuse multimodal information. While there are minor differences among these approaches, their method of fusing multimodal information in SSM is fundamentally similar. For instance, considering the RGB modality, given the RGB search tokens $H_n^{R,x}$ in the n-th layer, their approach is defined as follows:

$$X_n^{R,x} = \mathcal{L}(\mathcal{N}(H_n^{R,x})), Z_n^{R,x} = \sigma(\mathcal{L}(\mathcal{N}(H_n^{R,x}))), \quad (1)$$

$$Y'^{,R} = \mathcal{SSM}_f(\psi(X_n^{R,x})) + \mathcal{SSM}_b(\psi(X_n^{R,x})), \quad (2)$$

$$Y_n^{R,x} = \mathcal{L}(\mathcal{N}(Y'^{,R} \odot Z_n^{R,x} + Y'^{,E} \odot Z_n^{R,x})) + H_n^{R,x}. \quad (3)$$

Unlike our approach, they use modality-specific gating signals $Z_n^{R,x}$ to guide the interaction between modalities and then fuse them additively. However, this additive fusion scheme overlooks the significant modality information gap, leading to additional noise for modality-specific features. As illustrated in Figure 4, we conducted additional experiments by designing a CMS variant of such scheme. Table 1 B shows that this scheme results in a drop in Precision Rate (PR) and Success Rate (SR) by 1.2% and 1.6%, respectively, thereby highlighting the advantages of our approach. Furthermore, we explore two commonly used fusion architectures in multimodal vision tasks to explain how our fusion mechanism is well-suited for the RGBE tracking task. One approach is the Transformer-based Cross Attention used in CrossViT [2], and the other is the MLP-based efficient bidirectional adapter proposed in BAT [1]. As shown in Table 1 (C and D), the lack of a selective token interaction scheme in these approaches, fails to bridge the significant distribution gaps between different modalities, resulting in a drop in tracking performance.



(a) Illumination Variation (b) Background Influence (c) Out-of-View

Figure 2. Ground truth of event data (left), ground truth of RGB data (middle), and the MamTrack score map (right) on the FELT dataset under three long-term challenging conditions.

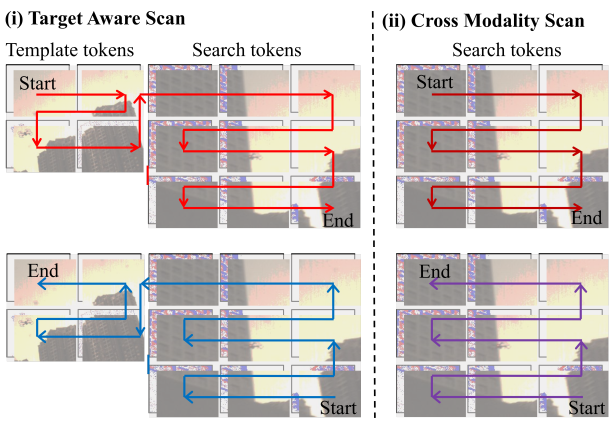


Figure 3. Visualization of the scanning scheme in different modules of FusionMamba: (i) Forward (top) and backward (bottom) scanning in the TAS module, and (ii) Forward (top) and backward (bottom) scanning in the CMS module.

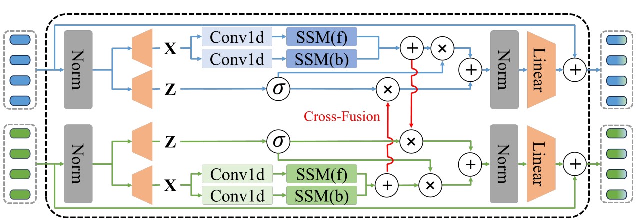


Figure 4. Variant of our Cross-Modality Scan module

References

[1] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 927–935, 2024. 1

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 1

[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1

[4] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv:2402.12192*, 2024. 1

[5] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-fetrack: Frame-event tracking via state space model. *arXiv preprint arXiv:2404.18174*, 2024.

[6] Haoyuan Li, Qi Hu, You Yao, Kailun Yang, and Peng Chen. CFMW: Cross-modality Fusion Mamba for Multispectral Object Detection under Adverse Weather Conditions, 2024. 1

[7] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model, 2024. 1