Exploiting Multimodal Spatial-temporal Patterns for Video Object Tracking

Xiantao Hu[1], Ying Tai[2,1*], Xu Zhao[1], Chen Zhao[2], Zhenyu Zhang[2], Jun Li[1], Bineng Zhong[3], Jian Yang[1*]

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Nanjing University
[3]Guangxi Normal University
{xiantaohu, csjyang, junli}@njust.edu.cn, {yingtai,zhenyuzhang}@nju.edu.cn, bnzhong@gxnu.edu.cn

Abstract

Multimodal tracking has garnered widespread attention as a result of its ability to effectively address the inherent limitations of traditional RGB tracking. However, existing multimodal trackers mainly focus on the fusion and enhancement of spatial features or merely leverage the sparse temporal relationships between video frames. These approaches do not fully exploit the temporal correlations in multimodal videos, making it difficult to capture the dynamic changes and motion information of targets in complex scenarios. To alleviate this problem, we propose a unified multimodal spatial-temporal tracking approach named STTrack. In contrast to previous paradigms that solely relied on updating reference information, we introduced a temporal state generator (TSG) that continuously generates a sequence of tokens containing multimodal temporal information. These temporal information tokens are used to guide the localization of the target in the next time state, establish long-range contextual relationships between video frames, and capture the temporal trajectory of the target. Furthermore, at the spatial level, we introduced the mamba fusion and background suppression interactive (BSI) modules. These modules establish a dual-stage mechanism for coordinating information interaction and fusion between modalities. Extensive comparisons on five benchmark datasets illustrate that STTrack achieves state-of-the-art performance across various multimodal tracking scenarios.

Code — https://github.com/NJU-PCALab/STTrack

Introduction

Visual object tracking is the process of locating and following a specific object across consecutive frames in a video sequence. As a fundamental vision task, it is essential for various applications (Wang et al. 2022, 2025; Jiang et al. 2023; Zhang et al. 2024b) and their related tasks (An et al. 2024; Zheng et al. 2023; Zhang et al. 2024a; Fang et al. 2023; Nan et al. 2024; Ning et al. 2023). Despite numerous efficient RGB-based trackers (Xie et al. 2024; Shi et al. 2024; Hu et al. 2024b; Wei et al. 2023; Chen et al. 2020, 2021, 2022; Xue et al. 2024) have been proposed through high quality dataset (Fan et al. 2019; Huang, Zhao, and Huang 2021; Müller et al. 2018), they are still limited by the degradation of RGB imaging quality caused by the complexity of real-world scenarios, which leads to tracking errors. Compared to RGB modalities, thermal infrared (TIR) provides clear target information in low light environments; depth modalities offer distance cues from depth cameras; and event modalities use event-based cameras to capture motion information and generate stable target trajectories. Therefore, developing an effective multimodal tracker that combines various modality X (such as TIR, depth, and event) with RGB is crucial for robust tracking.

Multimodal tracking methods can be broadly categorized

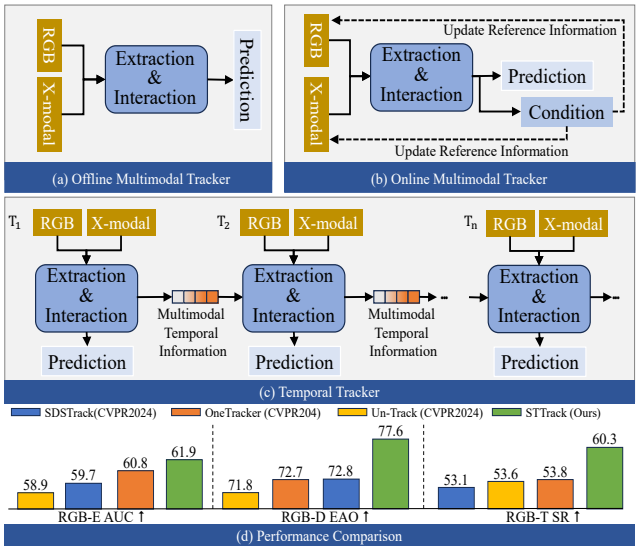

Figure 1: **Illustrations of different frameworks of multimodal trackers (a)-(c), and performance comparison (d)**. (a) Offline multimodal tracker performs offline tracking of video sequences using fixed template frames. (b) Online multimodal tracker is based on an updating strategy, which utilizes the results condition to update the reference information.(c) Our proposed STTrack transmits multimodal temporal information throughout the tracking process. (d) STTrack achieves superior performance against recent state-of-the-art competitors on three popular multimodal tasks.

*Ying Tai and Jian Yang are the corresponding authors.

into: *Offline trackers with fixed template frames* and *online trackers that update reference information*. 1) Traditional offline multimodal trackers focus on the fusion and interaction of spatial multimodal features, evolving from early CNN architectures (Yang et al. 2022; Xiao et al. 2022; Zhang et al. 2019) to the recent Transformer architectures (Chen et al. 2024; Hu et al. 2024a). As shown in Fig. 1 (a), offline trackers rely on a fixed initial target appearance as reference information for the entire tracking process. However, as time passes, the target may deform or become occluded, rendering the initial template frame unable to accurately capture its current state. 2) In contrast, as depicted in Fig. 1 (b), online trackers capture more recent target appearance features by updating reference information, such as template images (Wang et al. 2024; Luo et al. 2023), search images (Tang, Xu, and Wu 2022), or historical frame features (Zhang et al. 2023). Although these approaches enable updates at specific points in time, their reliance on sparse temporal relationships (*i.e.*, updates limited to specific conditions) neglects the continuity of temporal information. In video tracking tasks, *target changes and movements* typically follow a certain trend, which is challenging to capture and express without explicit temporal modeling, thus limiting the model's performance in complex scenarios.

To address this issue, we propose a novel tracking framework STTrack based on multimodal spatial-temporal patterns. STTrack improves to capture and represent the dynamic target by explicitly leveraging the temporal context within multimodal video data. As shown in Fig. 1 (c), we make full use of the multimodal temporal information from videos to guide the modeling of the current state of targets, thereby constructing a unified multimodal temporal strategy. There are several critical modules in STTrack. 1) At the temporal level, we design a novel *temporal state generator (TSG) based on cross mamba architecture* (Wan et al. 2024). TSG combines the current cross-modal target representation features with previous multimodal temporal information, employing an autoregressive mechanism to generate multimodal temporal information tokens for the current time step. These tokens act as bridges for information transfer, facilitating the tracking process for the next time node. 2) At the spatial level, since cross-modal interaction and fusion are crucial for effective multimodal tracking, we therefore propose the *background suppression interactive* module in the feature extraction stage of the visual encoder, and the *mamba fusion* module in the final modality fusion stage, respectively. The BSI module improves each modality branch's representation by integrating features from other modalities, while the mamba fusion module dynamically merges multimodal features from both branches to facilitate precise object localization.

We summarize the contributions of STTrack as follows:

- To fully exploit the temporal information from multiple modalities, we propose STTrack, which introduces temporal state generator to reveal temporal context of target.
- We propose the BSI and mamba fusion modules, which optimize information interaction and dynamic fusion between modalities during the feature extraction and modality fusion stages, respectively.
- The proposed STTrack achieves state-of-the-art performance on five popular multimodal tracking benchmarks, including RGBT234, LasHeR, VisEvEnt, Depthtrack, and VOT-RGBD2022.

Related Works

Multimodal Tracking. In recent years, multimodal tracking has gained widespread attention for its ability to achieve robust tracking in complex scenarios. By allowing different modalities to complement each other, it overcomes challenges that a single modality cannot address on its own. Early multimodal tracking methods (Zhang et al. 2021; Li et al. 2020) typically focused on specific multimodal task. For instance, APFNet introduces the concept of attribute fusion based on ResNet (He et al. 2016), enhancing its performance under specific challenges. TBSI (Hui et al. 2023) extends ViT (Dosovitskiy et al. 2021) to RGB-T tasks and leverages the TBSI module to optimize cross-modal interactions. More recently, some works (Hong et al. 2024) have begun exploring unified architectures capable of handling multiple multimodal tasks. ViPT (Zhu et al. 2023) integrates other modalities into the RGB modality through a prompt mechanism. SDSTrack (Hou et al. 2024) and BAT (Cao et al. 2024) explore symmetrical architectures for primary and auxiliary modality transformations. However, existing unified multimodal tracking frameworks often perform coarse multi-level interactions on all modality features within the encoder, inevitably introducing *irrelevant background noise into the search area*. In this work, we propose a novel BSI module to leverage the correlation strength among the template, temporal information, and search area to *emphasize target features while suppressing background interference*.

Temperoal Modeling. Temporal information is crucial for tracking models to capture long-term changes and motion trends of the target. In RGB-based object tracking, researchers (Yan et al. 2021; Chen et al. 2023; Lin et al. 2022; Song et al. 2023; Cui et al. 2022) have carefully designed various update strategies, typically guiding current state tracking through the fusion of accumulated templates. STMTrack (Fu et al. 2021) proposed a spatial-temporal memory network to exploit historical information. In contrast, multimodal tracking scenarios are more complex, requiring the consideration of not only RGB information but also the integration of additional modalities. In RGB-T Tracker, DMSTM (Zhang et al. 2023) uses a dual-modality space-time memory network to aggregate historical information as well as the apparent information of the current frame. TATrack (Wang et al. 2024) have successfully improved performance by combining dynamic template frames of the two modalities.

However, exploration in the temporal dimension of multimodal tracking currently faces two main challenges: 1) Updating reference materials, such as template images and historical frame information, often depends on preset conditions, leading to *sparse temporal relationships*. This limitation disrupts the continuous flow of information, making it difficult for the model to accurately capture ongoing target

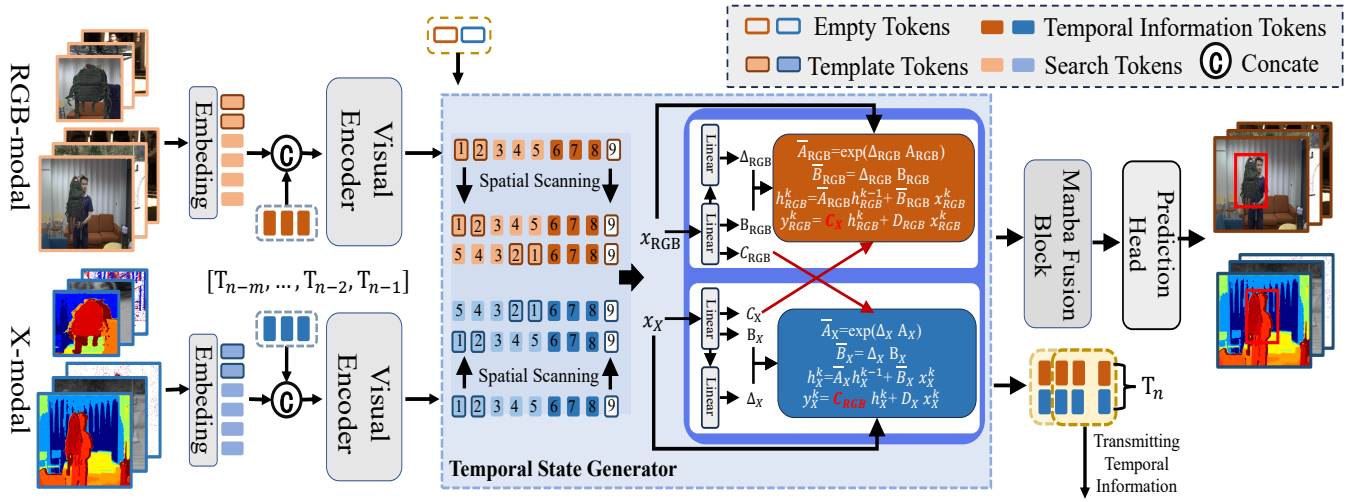

Figure 2: **Overall architecture of STTrack**. The temporal information tokens of each modality, along with the image tokens, are fed into the vision encoder to guide the extraction of current features using temporal information. In our designed Temporal State Generator, the current temporal tokens are generated based on cross-modal features and previous temporal features. We have added cross modal interaction in Visual Encode. Finally, the features are finely adjusted and fused through the mamba fusion module and then fed into the tracking head to predict the current state.

movements and changes over time. 2) Existing temporal exploration designs primarily focus on *single multimodal task*, limiting their effectiveness in multi-task environments. In contrast, our STTrack framework leverages explicit frame-to-frame temporal information exchange, capturing the target's temporal evolution using video context. Our method improves *temporal continuity and contextual coherence*, as verified in the experiment section, and demonstrates its potential for unified application across *various visual multimodal tasks*.

Methodology

In this paper, we introduce a novel spatial-temporal tracker (STTrack) based on temporal information, enabling continuous frame-to-frame information transfer through spatial-temporal data. Fig. 2 illustrates the overall architecture of STTrack. In this section, we first briefly review the state space model. Subsequently, we provide a detailed introduction to the overall architecture of our STTrack.

Preliminaries

The state space models draw inspiration from continuous linear time-invariant (LTI) systems. The aim of SSM is to transform a one-dimensional function or sequence, represented as $x(t)$, into $y(t)$ via the hidden space $h(t) \in R^N$ with linear complexity. The system can be represented mathematically by the following formula:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t), \end{aligned} \tag{1}$$

where the system's count parameters include the evolution parameter $A \in \mathbb{R}^{N\times N}$, projection parameters $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{1\times N}$, and skip connection $D \in \mathbb{R}$. The $h'(t)$ refers to the time derivative of $h(t)$, and N is the state size.

When handling discrete sequences such as images and text, state space models need to convert continuous-time signals into discrete-time signals to accommodate the nature of discrete data. SSM adopt zero-order hold (ZOH) discretization to map the input sequence $\{x^1, x^2, ..., x^k\}$ to the output sequence $\{y^1, y^2, ..., y^k\}$. Specifically, suppose Δ as the pre-defined timescale parameter to transformer continuous parameters A, B to discrete space $\overline{A}$, $\overline{B}$. The discretization process is defined as follows:

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

After the discreization, Eq. (1) can be rewritten as:

$$\begin{aligned} h^k &= \overline{A}h^{k-1} + \overline{B}x^k, \\ y^k &= Ch^k + Dx^k. \end{aligned} \tag{3}$$

SSM excels at modeling discrete sequences, but their inherent LTI property results in fixed parameters, making them insensitive to input variations. To overcome this limitation, a novel approach called the Selective State Space Model, also referred to as Mamba (Gu and Dao 2023; Li et al. 2024; He et al. 2024; Patro and Agneeswaran 2024; He et al. 2024), has been introduced. Mamba makes model parameters dependent on the input data. It derives matrices B, C, and Δ directly from the input x, allowing the model to adapt to different contexts and capture complex interactions within long sequences.

Tracking Process

The multimodal tracking task generally involves integrating two distinct video modalities, which collaboratively contribute to the final decision-making process for tracking objects. For the input, each modality's data is first converted

into the corresponding template tokens (Z_{RGB}, Z_X) and search tokens (S_{RGB}, S_X) through patch embedding and positional embedding encoding. These tokens are then concatenated with the temporal information tokens that generated from the previous time state and fed into the tracker together. As shown in Fig. 1 (c), STTrack constructs a bridge between spatial and temporal information through its architecture, which consists of *a visual encoder, a temporal state generator, a mamba fusion module, and a prediction head*. We employ ViT (Dosovitskiy et al. 2021) as the visual encoder, with shared weights across the encoders, and insert background suppression interactive modules after each transformer layer. The visual encoder dynamically extracts precise multimodal features from the input multimodal images and prior temporal information. These features are fed into the temporal state generator to produce the current temporal information tokens, which are then passed to the next time point. The tracker then refines and fuses the visual features in the mamba fusion module, ultimately delivering them to the prediction head for the final tracking results.

Temporal State Generator

Previous methods typically focused on multimodal spatial features to achieve precise tracking results. However, these trackers are less effective in addressing challenges such as *changes in moving targets and interference from similar objects*. To better capture target changes, it is crucial to construct stable inter-frame information features. We introduce a temporal state generator that merges the unidirectional recurrent approach of cross mamba, employing autoregressive modeling to seamlessly transfer information from previous time nodes to the current one. This process integrates current multimodal spatial information to generate the temporal information tokens T^{cur}_{RGB} and T^{cur}_{X}. Specifically, the temporal state generator takes features from two modalities $x_{RGB} = [Z_{RGB}; S_{RGB}; T^{pre}_{RGB}]$ and $x_X = [Z_X; S_X; T^{pre}_X]$ as input, generating the target state for the current time node and multimodal features after modality interaction. Where T^{pre} is the temporal information learned from the previous m frames. Notably, at this stage, an empty token as T^{cur} was inserted at the end to store the target information at the current time node. We first apply 1D convolution to x_{RGB} and x_X, then linearly project them to produce the features $\overline{B}$, $\overline{C}$, and D as described in the preliminaries. By exchanging the C matrix, the temporal state generator can incorporate complementary information from another modality when generating the current temporal token. Specifically, this process can be represented as:

$$\begin{aligned} h^k_{RGB} &= \overline{A}_{RGB} h^{k-1}_{RGB} + \overline{B}_{RGB} x^k_{RGB}, \\ y^k_{RGB} &= \boldsymbol{C_X} h^k_{RGB} + D_{RGB} x^k_{RGB}. \end{aligned} \tag{4}$$

$$\begin{aligned} h^k_{X} &= \overline{A}_{X} h^{k-1}_{X} + \overline{B}_{X} x^k_{X}, \\ y^k_{X} &= \boldsymbol{C_{RGB}} h^k_{X} + D_X x^k_X, \end{aligned} \tag{5}$$

where $k \in [1, 2, .., l]$, l is the length of visual tokens, and y^k_{RGB}, y^k_X are concatenated to generate the visual features y_{RGB}, y_X. The original mamba block is designed for the 1D sequence, which limits their ability to understand visual tokens with spatial location information. Therefore, we adopt the commonly used bidirectional scanning method (Zhu et al. 2024) in visual Mamba to process the visual tokens. Specifically, we reverse the order of the visual tokens and perform calculations, then add the results of the reversed calculations to those of the non-reversed calculations.

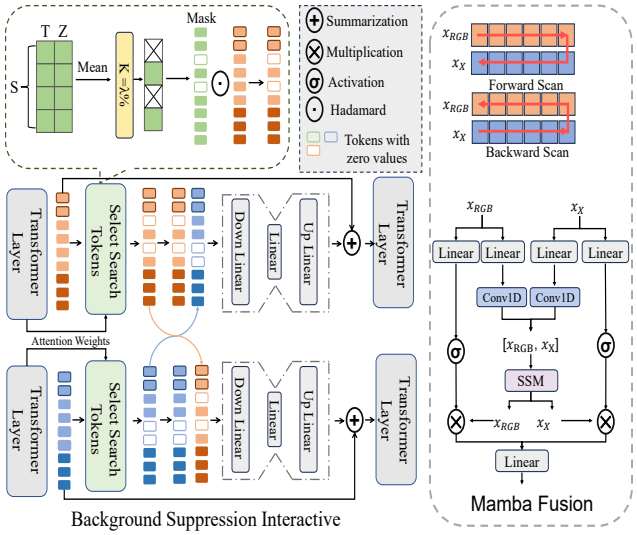

Figure 3: **Left**: Architecture of the background suppression interactive module. **Right**: Details of the fusion mamba. In BSI module S is a search areas tokens, Z denotes the template tokens and T is the temporal information tokens.

Extracting the current state information using the temporal information token allows us to add it to the queue T and propagate it to the next frame:

$$T = \begin{cases} [T_1, .., T_{t-1}, T_t] & \text{if } t < m \\ [T_{t-m}, .., T_{t-1}, T_t] & \text{if } t >= m. \end{cases} \tag{6}$$

where m is number of temporal information tokens, and t is time node. Temporal information tokens act as a bridge, linking the past, present, and future by using previous data to guide future modeling. Replacing the C matrix with cross-modal attention allows our temporal tokens to capture more comprehensive information.

Background Suppression Interactive

Incorporating multiple interactions within the encoder has become the mainstream method in multimodal tracking. To this end, we incorperate our background suppression interactive (BSI) module to each layer of the encoder to enhance cross-modal interactions. Our visual encoder retains ViT architecture, where its self-attention (Vaswani et al. 2017) mechanism is generally regarded as spatial aggregation of normalized tokens. Therefore, the similarity between tokens can be captured by the attention map, calculated as follows:

$$\begin{aligned} W_Z &= \text{softmax}(\frac{Q_Z \times K_S}{\sqrt{d}}), \\ W_T &= \text{softmax}(\frac{Q_T \times K_S}{\sqrt{d}}), \end{aligned} \tag{7}$$

where W_Z represents the correlation between the search area and the template features, while W_T represents the correlation between the search area and temporal information.

We use the attention computed in ViT as the criterion for background suppression, thereby avoiding additional computations. When calculating similarity with the search area, we use a 3×3 matrix centered on the template along with temporal information tokens to compute and average the results. Since tracking is essentially a matching task, a low similarity between search area tokens and the template region likely indicates a background area. Our temporal information tokens provide sufficient guidance to represent the target. By setting the filtering ratio λ, we first sort the search tokens by their similarity. *Then, we select the bottom λ proportion of tokens with the lowest similarity, mark them as invalid, and set their values to zero.*

Moreover, with each iteration of feature modeling by the visual encoder, the accuracy of the generated association matrix improves progressively. Therefore, we divide the filtration ratio of the 12 layer BSI into three parts and gradually increase the λ. As in Fig. 3, after background suppression, we concatenate the features from the two modalities and generate cross-modal feature prompts through linear layers:

$$\begin{aligned} x_{RGB}^{i} &= F_{RGB}^{i}([f_{RGB}; f_X]), \\ x_X^{i} &= F_X^{i}([f_X; f_{RGB}]), \end{aligned} \tag{8}$$

where f_{RGB}, f_X represent the features after background suppression, and i denotes the transformer layer number.

Mamba Fusion

Mamba excels in long-sequence modeling capabilities. Building on this, we concatenate the sequences of the two modalities and use a bidirectional scanning strategy to capture long-range dependencies in both modalities. Finally, we sum the two modality sequences to complete the modality fusion. The process can be represented as:

$$x = Fusion([Z_{RGB}, S_{RGB}], [Z_X, S_X]). \tag{9}$$

After obtaining $x_{\text{RGB}} \in R^{N \times C}$ and $x_X \in R^{N \times C}$, we concatenate them along the channel dimension and use a linear layer to adjust the dimension to C. Here N represents the number of tokens of the feature sequence and C represents the channel dimension. Detailed architecture is shown in Fig. 3. In this way, we refine and fuse the features before they are fed into the prediction head.

Head and Objective Loss

Following most of the latest multimodal tracking methods (Zhu et al. 2023; Wu et al. 2024), we employ a stacked set of Fully Convolutional Networks (FCNs) (Ye et al. 2022) to construct the prediction head. Notably, during the tracking process, we maintain a temporal tokens T with a length of m. We use L_{cls} (Lin et al. 2017) to denote the weighted focal loss for classification. For bounding box regression, we adopt the generalized IoU loss L_{iou} (Rezatofighi et al. 2019) and the L_1 loss. The overall loss function of STTrack is:

$$L = L_{\text{cls}} + \alpha L_{\text{iou}} + \beta L_1, \tag{10}$$

where $\alpha = 2$ and $\beta = 5$, which are hyperparameters to balance the contributions of loss terms.

Method	Source	LasHeR		RGBT234	
		SR ↑	PR ↑	MSR ↑	MPR ↑
STTrack	Ours	**60.3**	**76.0**	**66.7**	**89.8**
GMMT	AAAI'24	56.6	70.7	64.7	87.9
BAT	AAAI'24	56.3	70.2	64.1	86.8
TBSI	CVPR'24	56.3	70.5	64.3	86.4
TATrack	AAAI'24	56.1	70.2	64.4	87.2
OneTracker	CVPR'24	53.8	67.2	64.2	85.7
Un-Track	CVPR'24	53.6	66.7	61.8	83.7
SDSTrack	CVPR'24	53.1	66.5	62.5	84.8
ViPT	CVPR'23	52.5	65.1	61.7	83.5
OSTrack	ECCV'22	41.2	52.5	54.9	72.9
ProTrack	MM'22	42.0	53.8	59.9	79.5
APFNet	AAAI'23	36.2	50.0	57.9	82.7

Table 1: Comparisons on **RGB-T tracking**.

Method	VOT-RGBD22			DepthTrack		
	EAO ↑	Acc. ↑	Rob. ↑	F-score ↑	Re ↑	Pr ↑
STTrack	**77.6**	**82.5**	**93.7**	**63.3**	**63.4**	**63.2**
SDSTrack	72.8	81.2	88.3	61.4	60.9	61.9
OneTracker	72.7	81.9	87.2	60.9	60.4	60.7
Un-Track	71.8	82.0	86.4	61.2	61.0	61.3
ViPT	72.1	81.5	87.1	59.4	59.6	59.2
SBT-RGBD	70.8	80.9	86.4	-	-	-
OSTrack	67.6	80.3	83.3	52.9	52.2	53.6
DET	65.7	76.0	84.5	53.2	50.6	56.0
ProTrack	65.1	80.1	80.2	57.8	57.3	58.3
SPT	65.1	79.8	85.1	57.8	53.8	52.7
STARK-RGBD	64.7	80.3	79.8	-	-	-
KeepTrack	60.6	75.3	73.9	-	-	-
ATCAIS	55.9	76.1	73.9	47.6	45.5	50.0

Table 2: Comparisons on **RGB-Depth tracking**.

Experiment

In this section, we begin by detailing the experimental training procedures and the inference process of the proposed STTrack. Following this, we compare STTrack against other leading methods using various benchmark datasets.

Implementation Details

Training. We train on multiple multimodal tasks, including LasHeR for RGB-T tracking, VisEvent for RGB-E tracking, and DepthTrack for RGB-D tracking. For input data, we use two 128×128 template images and one 256×256 search image. The training was conducted on four NVIDIA Tesla A6000 GPUs over 15 epochs, with each epoch consisting of 60,000 sample pairs and a batch size of 32. AdamW (Loshchilov and Hutter 2018) was employed as the optimizer, with an initial learning rate of $1e{-}5$ for the ViT backbone and $1e{-}4$ for other parameters. After 10 epochs, the learning rate was reduced by a factor of 10.

Inference. During inference, we maintain the same training setting, using two template frames (includes a fixed initial template frame and a dynamically updated template frame.). Temporal information is incrementally incorporated into the tracking process, frame by frame. The tracking speed, tested on a NVIDIA 4090 GPU, is approximately 35.5 FPS.

Comparison with State-of-the-Arts

LasHeR. The LasHeR represents a substantial RGB-T tracking dataset comprising 1224 aligned sequences. As

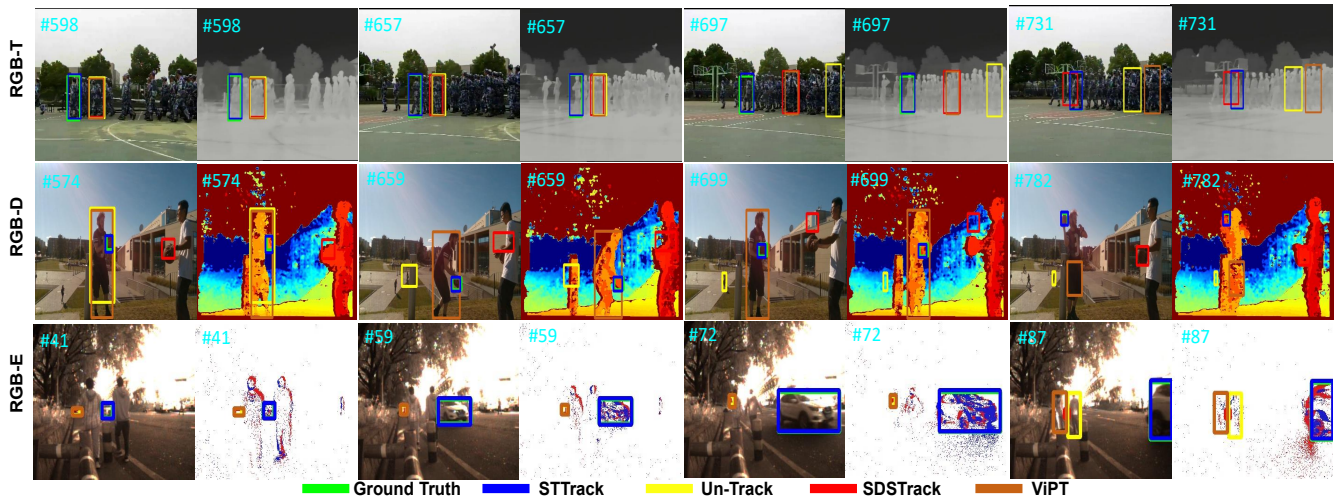

Figure 4: Qualitative comparison between our method and other unified multimodal trackers on three multimodal task. The three sequences correspond to scenarios involving similar object interference, fast motion, and target deformation. Our tracker effectively addresses these challenges through dual optimization in both the temporal and spatial dimensions.

	STARK_E	PrDiMP_E	LTMU_E	ProTrack	TransT_E	SiamRCNN_E	OSTrack	Un-Track	ViPT	SDSTrack	OneTrack	STTrack
AUC ↑	44.6	45.3	45.9	47.1	47.4	49.9	53.4	58.9	59.2	59.7	60.8	**61.9**
Pr ↑	61.2	64.4	65.9	63.2	65.0	65.9	69.5	75.5	75.8	76.7	76.7	**78.6**

Table 3: Comparisons on **RGB-Event tracking**.

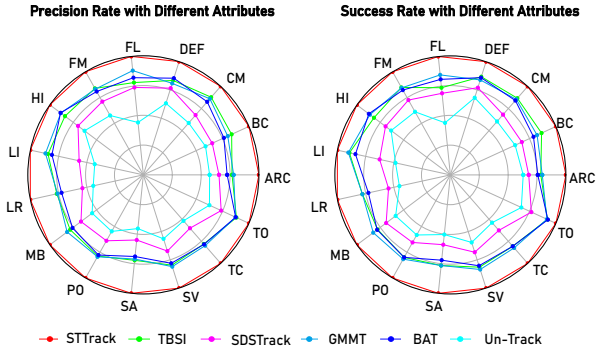

Figure 5: Comparison of STTrack and SOTA trackers (including unified trackers and RGB-T trackers) under different attributes in the LasHeR dataset.

shown in Tab. 1, our STTrack achieved an SR of 60.3% and a PR of 76.0%, surpassing GMMT by 4.3% in SR and 5.3% in PR, demonstrating the effectiveness of continuous spatial-temporal modeling.

RGBT234. The RGBT234 benchmark introduces enriched annotations and an expanded set of environmental challenges. As illustrated in Tab. 1, STTrack achieves the best MSR score of 66.7%, outperforming the recent trackers.

DepthTrack. DepthTrack is a long-time tracking dataset. The dataset covers 200 sequences, 40 scenes and 90 target objects. As shown in Tab. 2, our STTrack obtains SOTA results with 63.3% in F-score, 63.4% in recall, and 63.2% in precision.

#	Method	LasHeR	DepThTrack	Visevent	Δ
1	Baseline	56.0	58.8	60.0	–
2	+ Template Updata	57.1	59.5	59.8	**+0.5**
3	+ Temporal Information	58.9	62.0	61.1	**+1.8**
4	+ Mamba Fusion	59.2	62.1	61.3	**+0.2**
5	+ BSI Module	60.3	63.2	61.9	**+0.9**

Table 4: Quantitative comparison among different variants of STTrack on the LasHeR dataset, DepThTrack dataset and Visevent dataset. 'Δ' denotes the performance change (averaged over benchmarks) compared with previous variants.

VOT-RGBD2022. VOT-RGBD2022 comprises 127 brief RGB-D sequences and evaluates tracker performance using Accuracy, Robustness, and Expected Average Overlap (EAO) metrics. As illustrated in Tab. 2, our proposed tracker, STTrack, demonstrates a notable improvement in EAO, achieving a 4.8% increase compared to the previous SOTA tracker SDSTrack.

VisEvent. VisEvent is the largest RGB-E dataset, encompassing 500 training video sequences and 320 testing video sequences. As shown in Tab. 3, our STTrack obtains AUC and precession of 61.9% and 78.6%, respectively.

Ablation Studies

Component Analysis. In Tab. 4, we conducted an ablation study using the AUC in LasHeR, the Precession in DepThTrack, and the AUC in Visevent. The baseline used ViT as the visual encoder and fused the two modalities through

Dataset	1	2	4	8
LasHeR	59.6	59.9	60.3	60.0
DepthTrack	61.0	61.2	63.2	62.9
VisEvent	61.4	61.6	61.9	61.7
Δ	–	+0.2	+0.9	-0.3

Table 5: Ablation study on the number of temporal information tokens. We use gray color to denote our final trackers setting. 'Δ' denotes the performance change (averaged over benchmarks) compared with previous number setting.

#	Filtering Ratio (λ%)	LasHeR	DepThTrack	Visevent
1	[0%,0%,0%]	59.4	61.6	61.1
2	[15%,15%,15%]	59.8	62.9	51.7
3	[30%,30%,30%]	59.6	62.1	61.5
4	[0%,15%,30%]	60.3	63.2	61.9

Table 6: Ablation study on the ration in BSI module. We use gray color to denote our final trackers setting.

a convolutional layer before the prediction head. Through experimental results, we found that template updates can timely refresh the target's appearance information, compensating for the initial template's shortcomings. The model's performance was not optimal because of its sparse template update method and lack of frame-to-frame information transfer. Therefore, the introduction of temporal information led to a significant improvement, resulting in a 1.8% gain, which demonstrates the effectiveness of temporal information in addressing these issues. Furthermore, the results show that due to the differences in target representation across different modalities, optimizing the modality fusion process with mamba fusion module can further improve model performance. Additionally, by reducing interference from non-essential regions, our proposed background suppression scheme effectively enhances the performance of cross-modal interactions.

Number of Temporal Information Tokens. We investigate the impact of temporal information on the performance of STTrack, as shown in Tab. 5. As the number increases from 1 to 4, the model's performance improves, indicating the temporal information tokens positively contribute to optimization. However, when the number is increased further, performance declines, possibly due to earlier temporal information tokens failing to accurately describe the current target state, leading to the introduction of noise.

Filtering Ratio in BSI. To validate the impact of background suppression on performance, we conducted experiments with different background filtering ratios. Here, the 12-layer BSI module was processed in three stages. As Tab. 6 indicates that while a fixed background filtering ratio can enhance performance, a three-stage approach with progressively increased filtering ratios yields more significant improvements. This is because, within a single-stream structure, the features of the search area, guided by the template and temporal information tokens, need to progressively highlight the foreground target layer by layer.

Figure 6: The attention map of temporal information tokens with search area. These visual results are in LasHeR.

Exploration Study and Analysis

Attribute-based Performance. We analyze the performance of our method in various scenarios by evaluating it on different attributes of LasHeR dataset. As shown in Fig. 5, STTrack surpasses previous state-of-the-art trackers on these attributes. This improvement is due to our approach's enhanced temporal information and complementary spatial features, allowing STTrack to maintain stable tracking even when the target undergoes changes. STTrack excels in scenarios requiring temporal information, such as Partial Occlusion (PO) and Deformation (DEF), as well as in conditions with significant modality imaging differences, such as low-light and high-light situations.

Visualization Results. As shown in Fig. 4, we qualitatively compare STTrack with three other multimodal unified trackers. In the RGB-T sequences, where similar objects cause significant interference and the target has clear movement directions, STTrack leverages temporal information for stable tracking. In the RGB-D sequences, despite severe occlusion, our method captures the target by utilizing the complementary strengths of the RGB and Depth modalities along with continuous temporal information. In the RGB-E sequences, where the car moves at high speed and undergoes significant deformation due to changes in camera distance, STTrack effectively tracks the target by gradually adapting to these changes over time. Besides, we visualize the attention map of the temporal information tokens with search area, as shown in Fig. 6. It demonstrates that the continuous propagation of temporal markers and the focus on object temporal information can effectively capture and respond to the dynamic state of the target.

Conclusion

In this work, we propose a tracking framework named STTrack based on multimodal spatio-temporal patterns. By leveraging temporal context, STTrack effectively captures and represents dynamic targets. The tracker incorporates a temporal state generator to generate multimodal temporal information that supports the tracking process. Additionally, it is equipped with the BSI module and Mamba Fusion module, which optimize modality branch representation and fuse multimodal features at the spatial level. Compared to previous multimodal trackers, our approach achieves state-of-the-art performance across three multimodal tasks.

Acknowledgements

This work was funded by the National Science Fund of China, with Grant Nos. U24A20330, 62361166670, and 62406135, the Natural Science Foundation of Jiangsu Province under Grant No. BK20241198, and the AI & AI for Science Project of Nanjing University, Grant No. 14380007.

References

An, X.; Zhao, L.; Gong, C.; Wang, N.; Wang, D.; and Yang, J. 2024. SHaRPose: Sparse High-Resolution Representation for Human Pose Estimation. In *AAAI*, 691–699. AAAI Press.

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

Chen, L.; Zhong, B.; Liang, Q.; Zheng, Y.; Mo, Z.; and Song, S. 2024. Top-down Cross-modal Guidance for Robust RGB-T Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *CVPR*, 14572–14581. IEEE.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *CVPR*, 8126–8135.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese Box Adaptive Network for Visual Tracking. In *CVPR*, 6667–6676. Computer Vision Foundation / IEEE.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R.; Tang, Z.; and Li, X. 2022. SiamBAN: Target-aware tracking with Siamese box adaptive network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5158–5173.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In *CVPR*, 13598–13608. IEEE.

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 *ICLR*. OpenReview.net.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *CVPR*, 5374–5383. Computer Vision Foundation / IEEE.

Fang, W.; Zhang, G.; Zheng, Y.; and Chen, Y. 2023. Multi-Task Learning for UAV Aerial Object Detection in Foggy Weather Condition. *Remote Sensing*, 15(18): 4617.

Fu, Z.; Liu, Q.; Fu, Z.; and Wang, Y. 2021. STMTrack: Template-Free Visual Tracking With Space-Time Memory Networks. In *CVPR*, 13774–13783. Computer Vision Foundation / IEEE.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

He, X.; Cao, K.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2024. Pan-mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv:2402.12192*.

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 *CVPR*, 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 *CVPR*, 26551–26561.

Hu, X.; Zhong, B.; Liang, Q.; Zhang, S.; Li, N.; and Li, X. 2024a. Towards Modalities Correlation for RGB-T Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hu, X.; Zhong, B.; Liang, Q.; Zhang, S.; Li, N.; Li, X.; and Ji, R. 2024b. Transformer Tracking via Frequency Fusion. *IEEE Trans. Circuits Syst. Video Technol.*, 34(2): 1020–1031.

Huang, L.; Zhao, X.; and Huang, K. 2021. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5): 1562–1577.

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 *CVPR*, 13630–13639. IEEE.

Jiang, L.; Wang, C.; Ning, X.; and Yu, Z. 2023. LTTPoint: A MLP-Based Point Cloud Classification Method with Local Topology Transformation Module. In *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, 783–789. IEEE.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware RGBT tracking. In *European conference on computer vision*, 222–237. Springer.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. VideoMamba: State Space Model for Efficient Video Understanding. arXiv:2403.06977.

Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; and Ling, H. 2022. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35: 16743–16754.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*, 2999–3007. IEEE Computer Society.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Luo, Y.; Guo, X.; Feng, H.; and Ao, L. 2023. RGB-T tracking via multi-modal mutual prompt learning. *arXiv preprint arXiv:2308.16386*.

Müller, M.; Bibi, A.; Giancola, S.; Al-Subaihi, S.; and Ghanem, B. 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, 310–327. Springer.

Nan, K.; Xie, R.; Zhou, P.; Fan, T.; Yang, Z.; Chen, Z.; Li, X.; Yang, J.; and Tai, Y. 2024. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*.

Ning, E.; Zhang, C.; Wang, C.; Ning, X.; Chen, H.; and Bai, X. 2023. Pedestrian Re-ID based on feature consistency and contrast enhancement. *Displays*, 79: 102467.

Patro, B. N.; and Agneeswaran, V. S. 2024. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*, 658–666. Computer Vision Foundation / IEEE.

Shi, L.; Zhong, B.; Liang, Q.; Li, N.; Zhang, S.; and Li, X. 2024. Explicit Visual Prompts for Visual Object Tracking. In *AAAI*, 4838–4846. AAAI Press.

Song, Z.; Luo, R.; Yu, J.; Chen, Y. P.; and Yang, W. 2023. Compact Transformer Tracker with Correlative Masked Modeling. In *AAAI*, 2321–2329. AAAI Press.

Tang, Z.; Xu, T.; and Wu, X.-J. 2022. Temporal aggregation for adaptive rgbt tracking. *arXiv preprint arXiv:2201.08949*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, Z.; Wang, Y.; Yong, S.; Zhang, P.; Stepputtis, S.; Sycara, K.; and Xie, Y. 2024. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*.

Wang, C.; Ning, X.; Sun, L.; Zhang, L.; Li, W.; and Bai, X. 2022. Learning discriminative features by covering local geometric space for point cloud analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Wang, C.; Wu, M.; Lam, S.-K.; Ning, X.; Yu, S.; Wang, R.; Li, W.; and Srikanthan, T. 2025. Gpsformer: A global perception and local structure fitting-based transformer for point cloud understanding. In *European Conference on Computer Vision*, 75–92. Springer.

Wang, H.; Liu, X.; Li, Y.; Sun, M.; Yuan, D.; and Liu, J. 2024. Temporal Adaptive RGBT Tracking with Modality Prompt. In *AAAI*, 5436–5444. AAAI Press.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive Visual Tracking. In *CVPR*, 9697–9706. IEEE.

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 *CVPR*, 19156–19166.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-Based Progressive Fusion Network for RGBT Tracking. In *AAAI*, 2831–2838. AAAI Press.

Xie, J.; Zhong, B.; Mo, Z.; Zhang, S.; Shi, L.; Song, S.; and Ji, R. 2024. Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers. In *CVPR*, 19300–19309.

Xue, C.; Zhong, B.; Liang, Q.; Xia, H.; and Song, S. 2024. Unifying Motion and Appearance Cues for Visual Tracking via Shared Queries. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 10448–10457.

Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Prompting for Multi-Modal Tracking. In *ACM Multimedia*, 3492–3500. ACM.

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 *ECCV*, volume 13682 of *Lecture Notes in Computer Science*, 341–357. Springer.

Zhang, F.; Peng, H.; Yu, L.; Zhao, Y.; and Chen, B. 2023. Dual-modality space-time memory network for RGBT tracking. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–12.

Zhang, H.; Chen, S.; Luo, L.; and Yang, J. 2024a. Few-shot learning with long-tailed labels. *Pattern Recognition*, 156: 110806.

Zhang, H.; Ning, X.; Wang, C.; Ning, E.; and Li, L. 2024b. Deformation depth decoupling network for point cloud domain adaptation. *Neural Networks*, 180: 106626.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; and Khan, F. S. 2019. Multi-Modal Fusion for End-to-End RGB-T Tracking. In *ICCV Workshops*, 2252–2261. IEEE.

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.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and Du, Y. 2023. Curricular Contrastive Regularization for Physics-Aware Single Image Dehazing. In *CVPR*, 5785–5794. IEEE.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual Prompt Multi-Modal Tracking. In *CVPR*, 9516–9526. IEEE.

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. arXiv:2401.09417.