Motion-Aware Object Tracking via Motion and Geometry-Aware Cues

Hongtao Yang[1,2], Bineng Zhong[1,2*], Qihua Liang[1,2], Xiantao Hu[1,3], Yufei Tan[1,2,4], Haiying Xia[4], Shuxiang Song[4*]

[1]Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
[2]University Engineering Research Center of Educational Intelligent Technology, Guangxi Normal University, Guilin 541004, China
[3]Nanjing University of Science and Technology
[4]Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, School of Electronic and Information Engineering, Guangxi Normal University, Guilin 541004, China
yht@stu.gxnu.edu.cn, bnzhong@gxnu.edu.cn, qhliang@gxnu.edu.cn, xiantaohu@njust.edu.cn
jeffrey.yf.tan@gxnu.edu.cn, xhy22@mailbox.gxnu.edu.cn, songshuxiang@mailbox.gxnu.edu.cn

Abstract

Understanding motion is essential for visual object tracking, especially in complex and dynamic scenarios. Yet, many existing methods rely on simplistic strategies such as template updates or temporal feature propagation, often overlooking the deeper modeling of motion information. To mitigate this limitation, we introduce a motion-aware spatio-temporal framework that enhances motion perception by explicitly matching motion patterns and modeling inter-frame motion relationships. Central to our design is a motion pattern dictionary, which encodes a diverse set of representative motion cues as learnable features. During tracking, features from the search region interact with the dictionary to retrieve the most relevant motion patterns, allowing the model to adapt to the current motion state. A dedicated decoder further incorporates temporal correlations to refine motion awareness. To complement motion modeling, we embed geometric cues into the search region features, which strengthens spatial perception, reduces ambiguity under occlusion, and improves foreground-background separation. Extensive evaluations on seven challenging benchmarks demonstrate the effectiveness of our design. In particular, MoDTrack$_{384}$ surpasses recent SOTA trackers on LaSOT by 1.2% in AUC, highlighting the benefits of motion pattern modeling and geometry-guided enhancement in mitigating tracking drift.

Introduction

Visual object tracking (Bertinetto et al. 2016; Chen et al. 2022b; Hu et al. 2024b; Ge et al. 2024; Zheng et al. 2025) is a fundamental task in computer vision, with broad applications in mobile robotics (Pereira et al. 2022), video surveillance (Cheng, Wang, and Li 2022), and autonomous driving (Premachandra, Ueda, and Suzuki 2020). Its dynamic and long-term nature, characterized by changes in appearance, motion, and other factors, makes it particularly challenging. In this context, effectively capturing spatio-temporal information is essential for robust tracking performance.

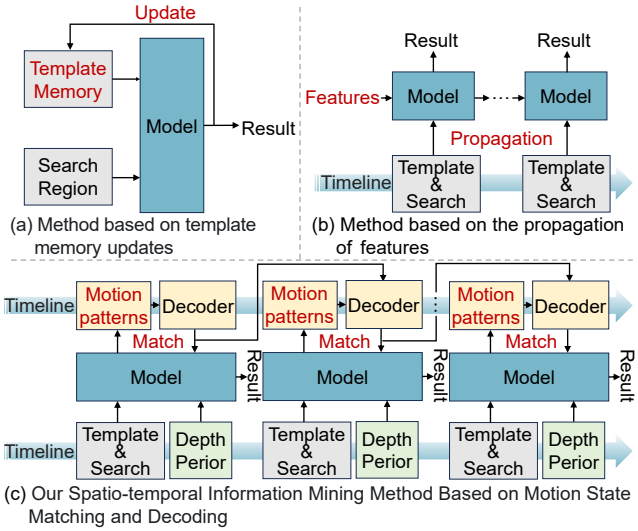

Figure 1: Comparison of different information mining methods. (a) Template memory update methods dynamically adjust templates based on tracking results. (b) Temporal feature propagation methods propagate appearance, trajectory, and other features along the time axis to assist in tracking. (c) Our method improves foreground-background separation by incorporating geometric cues, while continuously matching and decoding motion patterns to enhance the perception of target motion patterns.

Existing approaches utilize spatio-temporal information via two primary strategies. The first (Yan et al. 2021a; Cui et al. 2022), illustrated in Fig.1(a), relies on template memory updates, where dynamic templates adapt to target appearance changes. While this boosts adaptability, inaccurate updates can introduce tracking drift and performance degradation. The second (Zheng et al. 2024; Wei et al. 2023; Bai et al. 2024; Xie et al. 2024b; Hu et al. 2025a), shown in Fig.1(b), incorporates temporal feature propagation to enhance tracking. For instance, ODTrack (Zheng et al. 2024)

*Corresponding author.

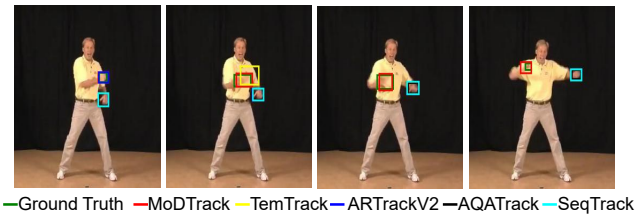

Figure 2: Comparison of MoDTrack with other state-of-the-art trackers in continuous motion scenarios.

propagates target features over time, compressing global motion information into a single token. ARTrack (Wei et al. 2023) predicts trajectories by propagating tracking coordinates across frames, while LMTrack (Xu et al. 2025) dynamically aggregates high-quality spatio-temporal features. However, these methods often fail to accurately capture rapid transitions in target motion patterns. When tracking objects with complex motion, their limited responsiveness to sudden changes causes drift, impairing tracking accuracy and stability. As shown in Fig.2, intricate hand motion leads several state-of-the-art trackers to gradually drift, losing the true target position and ultimately locking onto incorrect regions. In contrast, our method effectively mines and utilizes motion information, enabling better understanding of complex motion and preventing drift in such scenarios.

Studies in human vision (Chetverikov and Jehee 2023; Bill, Gershman, and Drugowitsch 2022) suggest that motion perception plays a vital role in robust tracking. However, motion estimation often suffers from uncertainty due to noise, occlusions, or abrupt changes in speed and direction. To address this challenge, the human brain goes beyond simple motion cues by integrating various motion-related priors, including velocity, motion streaks, and motion history. These priors are organized into structured representations that capture coherent motion and hierarchical patterns. Inspired by this mechanism, we propose a motion-aware spatio-temporal modeling framework that explicitly encodes and retrieves motion patterns for accurate tracking (Fig.1(c)). In our framework, motion patterns are represented as learnable features within a motion pattern dictionary, allowing the model to retrieve relevant motion information by matching features in the search region with the dictionary. The decoder further incorporates inter-frame motion relationships to enhance motion awareness. In parallel, we progressively introduce geometric cues to strengthen spatial perception, which helps alleviate ambiguity under occlusion and improves foreground-background separation. By jointly modeling motion patterns and geometric cues, the proposed framework alleviates the limitations of existing spatio-temporal approaches and enhances robustness in complex motion scenarios.

In summary, our contributions are as follows:

- We propose a motion-aware tracking framework that explicitly models motion patterns as learnable features. The motion pattern dictionary effectively captures the temporal evolution of motion, enabling the tracker to better adapt to dynamic and complex target behaviors.
- We incorporate geometric cues to enhance spatial awareness and improve foreground-background separation, aiding target distinction in cluttered or occluded scenes. Integrating geometric cues with motion pattern modeling ensures more reliable tracking, especially when motion information alone fails to resolve ambiguities.
- Extensive experiments on seven challenging benchmarks demonstrate the effectiveness of our approach. On the LaSOT, MoDTrack$_{384}$ achieves a 74.4% AUC score, outperforming several state-of-the-art trackers and highlighting the advantages of combining motion pattern modeling with depth priors for long-term tracking.

Related Work

Spatio-Temporal Tracking

Visual tracking has progressed with improved target representation and matching strategies. Traditional trackers (Bertinetto et al. 2016; Chen et al. 2022b; Zhang, Fu, and Zheng 2022) locate targets by matching template and search regions, with SiamFC (Bertinetto et al. 2016) pioneering Siamese networks to balance speed and accuracy. The rise of transformers led many trackers (Chen et al. 2021; Yan et al. 2021a; Ye et al. 2022; Hu et al. 2025b, 2024a; Ge et al. 2025; Wang, Li, and Ge 2025) to adopt transformer-based backbones for stronger features and better interactions. For instance, TransT (Chen et al. 2021) uses attention for template-search interaction, OSTrack (Ye et al. 2022) unifies feature extraction and modeling with ViT, and SimTrack (Chen et al. 2022a) focuses on key search areas via a foveal window. However, relying on static templates makes these methods vulnerable to occlusion, deformation, lighting changes, and clutter. To address this, recent methods leverage spatio-temporal cues: STARK (Yan et al. 2021a), CTTrack (Song et al. 2023), and SeqTrack (Chen et al. 2023) update templates dynamically, while ODTrack (Zheng et al. 2024), ARTrack (Wei et al. 2023), and AQATrack (Xie et al. 2024b) emphasize temporal modeling through feature propagation or sequence prediction. Specifically, ODTrack propagates tokens over time, ARTrack treats tracking as coordinate regression, and AQATrack combines static features with short-term motion priors via autoregressive queries.

While these methods improve adaptability by using past features like historical appearances, they lack thorough motion modeling. Our method introduces a motion-aware spatio-temporal framework that explicitly stores and retrieves learned motion patterns. Rather than relying only on initial appearance or past features, we use attention to match search features with the motion pattern dictionary. This enables the tracker to fully leverage matched motion patterns for better motion awareness. By combining historical and current motion patterns, our approach captures motion information more effectively than prior methods.

Leveraging Geometric Cues for Tracking

Purely visual object tracking often struggles in challenging scenarios such as heavy occlusion, illumination variations,

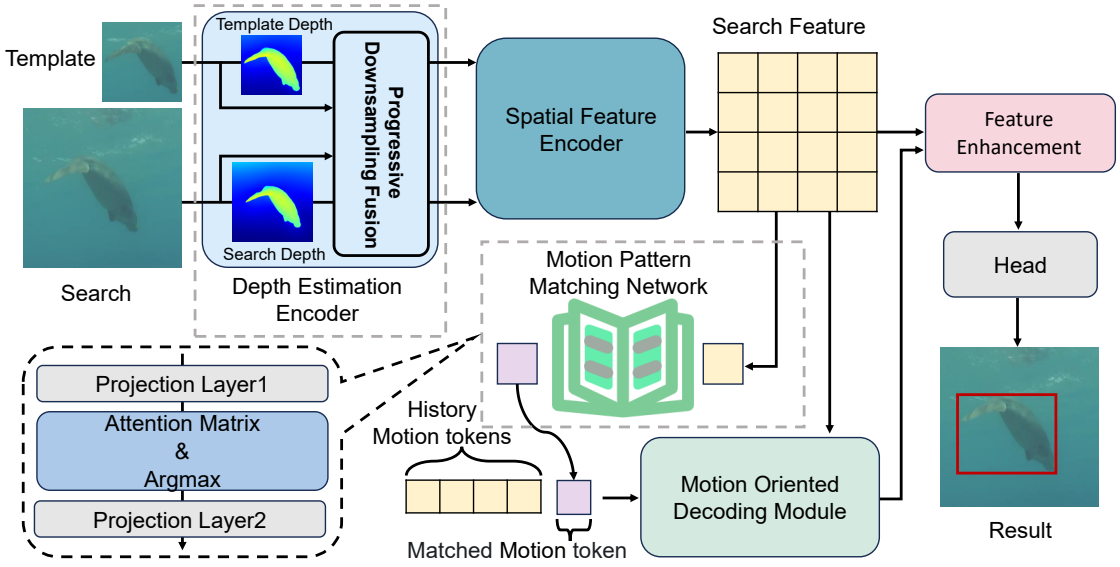

Figure 3: Overview of our framework. It consists of four main components: (1) a depth estimation encoder, which extracts depth features from the image and integrates them into spatial features; (2) a spatial feature encoder, which further refines the extracted features to obtain richer representations; (3) a matching network and an N-layer decoder, responsible for estimating the target's motion pattern by incorporating historical tokens and performing cross-attention; (4) a feature enhancement module, which refines the features before feeding them into the prediction head, producing the final tracking result.

and complex backgrounds, where reliance solely on appearance information may lead to tracking failures. To address this limitation, an increasing number of studies (Kart, Kamarainen, and Matas 2018; Qian et al. 2021; Wang et al. 2025) have explored incorporating additional geometric cues as a complementary source of visual information. Geometric priors provide valuable spatial context, helping distinguish objects from the background and improving performance in ambiguous or cluttered environments. For instance, DAL (Qian et al. 2021) designed a geometric-aware deep correlation filter that effectively utilizes geometric cues to enhance target localization. Yan et al. proposed DeT (Yan et al. 2021b), a trainable RGB-D tracker equipped with two-stream feature extraction backbones, enabling joint learning of geometric and appearance features. DMTrack (Gao et al. 2022) introduces a dual-fusion modality-aware tracker designed to learn both target information and discriminative representations. However, most of these methods rely heavily on large-scale RGB-D datasets to learn appearance-guided geometric representations for target localization, often overlooking the importance of motion modeling and the potential of geometric cues in refining motion patterns. This oversight ultimately leads to suboptimal performance in complex and dynamic motion scenarios.

In contrast, our method generates geometric cues without relying on large-scale RGB-D datasets or additional manual annotations. To reduce computational complexity, we seamlessly integrate geometric priors as auxiliary information into the original RGB image features, significantly enhancing their representation and robustness. This integration improves spatial awareness, effectively alleviates ambiguities caused by occlusion, and reduces biases in motion awareness through better foreground-background separation.

Methodology

Overview

The proposed tracker consists of three main components: a depth estimation encoder, a spatial feature encoder, and a motion pattern matching and decoding module. Given a template image $\mathbf{Z} \in \mathbb{R}^{H_z \times W_z \times 3}$ and a search image $\mathbf{X} \in \mathbb{R}^{H_x \times W_x \times 3}$, the depth estimation encoder generates corresponding depth maps $\mathbf{Z}_d$ and $\mathbf{X}_d$. During this process, depth features are progressively fused into RGB features as spatial priors. The spatial feature encoder then extracts enriched representations using attention to emphasize target-relevant features. The motion pattern module operates in two stages. First, search features are matched with stored motion patterns to identify the current motion state. Then, the selected pattern is decoded along with propagated temporal information and current features to model motion dynamics. Finally, the tracker outputs predictions via a center-head network, following prior designs (Ye et al. 2022; Xie et al. 2024b).

Depth Estimation Encoder

The geometric cue encoder provides additional valuable visual knowledge to address challenges in complex and variable tracking scenarios. Geometric cues are particularly useful in dynamic real-world environments where appearance alone may be insufficient, such as when objects are partially occluded or undergo rapid motion. Inspired by Depth Anything V2 (Yang et al. 2024), we first extract image features using the dinov2 encoder. From this encoder, we obtain outputs from specific layers, denoted as f_t for layer

$t \in \{1, 2, \ldots, m\}$. These features represent different levels of abstraction of the image and capture important semantic and spatial information. To further process these features, we pass them through convolutional layers, which generate multi-scale representations that capture fine-grained details. These representations are then adaptively fused to create a single-channel geometric map, which will later be used to enrich the tracking model with critical spatial information.

To effectively leverage the visual knowledge provided by geometric maps, we employ a step-by-step downsampling approach for incorporating geometric priors into the tracking model. The process begins with the template and search images, which are passed through a 4×4 embedding layer. This transforms both images into patches with 128 channels, ensuring that the input data is appropriately shaped for subsequent processing. Similarly, the geometric information undergoes the same embedding operation to generate corresponding patches, aligning geometric and RGB features for later fusion. These features are progressively refined through a Multi-Layer Perceptron (MLP) layer and a downsampling layer, which reduces spatial dimensions while preserving crucial information. During this process, the geometric priors are adaptively adjusted through a feature adapter, as shown in Fig.4. The structure of the feature adapter is like an hourglass, allowing it to adjust the feature dimensions before fusing and enhancing them with the original image features. By fusing geometric priors with image features, we enhance the overall feature representation, allowing the model to utilize both appearance and spatial knowledge. This enriched representation is then passed to the spatial feature encoder, ultimately improving tracking performance, especially in challenging scenarios with occlusions, deformations, or sudden appearance changes.

Spatial Feature Encoder

The spatial feature encoder extracts features from the input image, focusing on the target's spatial representation. Unlike previous approaches (Ye et al. 2022; Zheng et al. 2024; Wei et al. 2023) that typically use Vision Transformers (ViT) for spatial encoding, our method adopts the Fast-iTPN (Tian et al. 2024) encoder, as in TemTrack (Xie et al. 2024a). This choice balances efficiency with the ability to capture fine-grained details from both the template and search regions.

In our approach, the template and search tokens, along with position and depth features, are concatenated into a single input, denoted as F_{zx}^0. This input is passed through an N-layer transformer encoder, where spatial features are progressively refined. By incorporating both appearance and geometric cues, the encoder can capture richer contextual information, leading to improved tracking performance. The output after each transformer layer is denoted as F_{zx}^l, and the refinement process is described as follows:

$$F_{zx}^l = \varphi^l(F_{zx}^{l-1}), \quad l = 1, 2, ..., N \tag{1}$$

Here, $F_z \in \mathbb{R}^{N_z \times D}$ and $F_x \in \mathbb{R}^{N_x \times D}$ represent 1D image token sequences, with $N_z = H_z \times W_z/16^2$, $N_x = H_x \times W_x/16^2$, and $D = 512$. Each transformer layer consists of multi-head attention and an MLP, with a learnable scaling factor that controls the influence of their outputs. F_{zx}^l denotes the output of the token sequence produced by the l-th transformer layer φ^l.

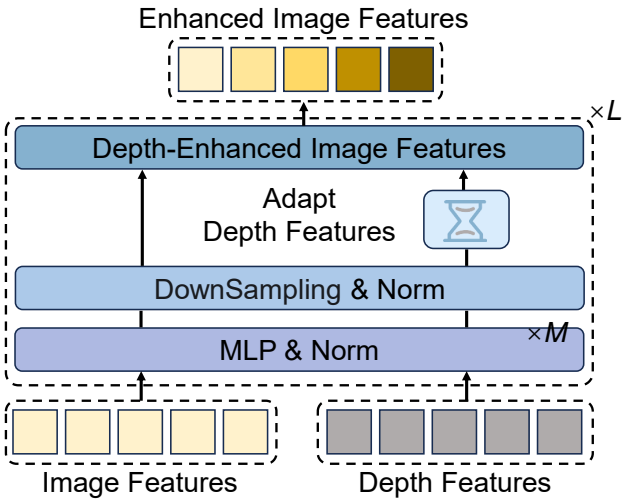

Figure 4: During downsampling, geometric cues serve as priors to enhance image features. After separate processing, the two are fused adaptively. This process iterates L times to produce the final geometry-enhanced features.

Motion Pattern Matching and Decoding

Motion perception plays a fundamental role in understanding motion information, and structured motion modeling facilitates more comprehensive motion awareness. During training, we learn the motion patterns of the object from video sequences, constructing a motion pattern dictionary that captures the diverse motion behaviors of the object. During inference, the current motion pattern of the object is retrieved through attention matrix matching, and the decoder integrates the motion relationships between frames to refine the tracking process. This approach allows the model to adapt effectively to motion variations, leveraging both historical and current motion information to enhance tracking accuracy in dynamic and complex environments.

We define a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{K \times C}$ as the motion pattern dictionary, where K is the number of embeddings and C is the feature dimension. The embeddings are initialized from a uniform distribution in $[-1/K, 1/K]$ to ensure diversity in the representation space. Given an input feature $\mathbf{F} \in \mathbb{R}^{B \times N \times C}$, where B is the batch size and N is the number of tokens, we perform the following steps:

Feature Projection. To improve representation and enable effective matching, the input features $\mathbf{F}$ are projected into the same space as the embedding matrix $\mathbf{E}$ via a learnable linear transformation.

Attention Matrix Matching. We compute an attention matrix $\mathbf{A} \in \mathbb{R}^{B \times N \times K}$ by applying dot-product similarity between the projected feature $\mathbf{F}'$ and the embedding matrix $\mathbf{E}$, where each entry in $\mathbf{A}$ measures the similarity between a feature token and each motion pattern embedding. For each token, we then identify the most relevant embed-

ding by selecting the one with the highest attention score. This matching mechanism effectively associates each token with its best-fitting motion pattern.

Encoding & Quantization. Once the most relevant embedding index is determined, the corresponding motion pattern embedding is retrieved to construct the quantized feature representation: $\mathbf{F}_q = \mathbf{E}[\text{index}]$. To further refine the quantized feature and enhance adaptability, we apply an additional projection layer:

$$\mathbf{F}'_q = \text{Proj}_2(\mathbf{F}_q), \tag{2}$$

here, $\text{Proj}_2(\cdot)$ is a fully connected layer that refines the retrieved motion pattern embeddings. The final quantized representation $\mathbf{F}'_q$ is then used for motion pattern decoding.

Motion Pattern Guided Decoding. Finally, the cross-attention module Φ_{CA} is utilized for feature interaction. In this process, the matched current motion patterns of the object are used as the query. First, in the first cross-attention operation, the historical motion information from previous frames acts as both the key and the value. The module calculates the matching relationship between the query and this historical motion information to integrate the motion relationships between frames. The output of this step, O_1, can be represented as: $O_1 = \Phi_{\text{CA}_1}(Q, K_h, V_h)$. Then, in the second cross-attention operation, the image features of the current frame serve as the key and the value. The module again computes the matching relationship between the updated query (from the first operation) and the current frame's features: $O_2 = \Phi_{\text{CA}_2}(O_1, K_c, V_c)$. This step helps to further refine the feature representation by integrating the latest information from the current frame. The overall purpose of the cross-attention module Φ_{CA} is to achieve better motion perception through temporal motion variations. This strategy improves motion perception through explicit modeling. Compared to prior methods, it better captures continuous motion changes, reducing drift in complex scenarios.

Prediction Head and Training Loss

We adopt a center-based prediction head to estimate the target's center and scale. Given feature X', the head outputs a center score map, size map, and offset map. The bounding box is computed by locating the peak in the score map and retrieving corresponding size and offset values, resulting in a normalized box $(x_{\text{center}}, y_{\text{center}}, \text{width}, \text{height})$. During the training process, we utilize Focal (Lin 2017) loss and GIoU (Rezatofighi et al. 2019) loss to regulate the model's learning. The total loss L can be expressed as:

$$L = L_{cls} + \lambda_1 L_{iou} + \lambda_2 L_1, \tag{3}$$

where λ_1 = 2 and λ_2 = 5 are the regularization parameters.

Experiments

Implementation Details

Our method is implemented in PyTorch and trained on four A800 GPUs. Speed is evaluated on a single V100. We report two TemTrack variants with different settings:

- MoDTrack$_{256}$. The resolution of template image and search region is 128×128 and 256×256 pixels.
- MoDTrack$_{384}$. The resolution of template image and search region is 192×192 and 384×384 pixels.

We use Depth Anything V2 (Yang et al. 2024) as the depth estimation encoder and adopt Fast-iTPN (Tian et al. 2024) as the spatial encoder, initializing it with the pretrained checkpoints of Fast-iTPN-B$_{224}$.

Training. We follow standard training protocols using COCO, GOT-10k, LaSOT, and TrackingNet datasets. For GOT-10k evaluation, the model is trained on the full training split. Data augmentations include brightness jittering and horizontal flip. MoDTrack is trained with AdamW optimizer, with backbone learning rate 4×10^{-5}, 4×10^{-4} for other parameters, and weight decay 10^{-4}, following OSTrack. Training lasts 150 epochs with 60k image pairs per epoch, and learning rate drops 10× after 120 epochs. For GOT-10k, training is 100 epochs with decay at epoch 80.

Inference. During inference, matched motion pattern features accumulate historical motion using a sliding window of size 4, consistent with training. Tab.1 compares model variants on parameters, FLOPs, and inference speed. MoDTrack$_{256}$ and MoDTrack$_{384}$ run on an NVIDIA V100 GPU, achieving 44 and 33 FPS, respectively.

Model	Source	Input Resolution	Params (M)	FLOPs (G)	Speed (fps)
MoDTrack$_{256}$	Ours	256×256	89	35.2	44
MoDTrack$_{384}$	Ours	384×384	89	79.1	33
SeqTrack$_{256}$	CVPR2023	256×256	89	66	44
SeqTrack$_{384}$	CVPR2023	384×384	89	148	15

Table 1: Comparison of our method and variants with other top trackers in parameters, complexity, and speed.

Comparison with the SOTA

GOT-10k (Huang, Zhao, and Huang 2019). GOT-10K is a large-scale benchmark for single object tracking, with over 10,000 video sequences covering a diverse range of object categories and challenging scenarios. Unlike conventional datasets, GOT-10K ensures a strict separation between the object classes in the training and test sets, promoting the evaluation of a tracker's generalization ability. As demonstrated in Tab.3, MoDTrack$_{384}$ achieves 76.3%, 86.4%, and 75.6% in metrics AO, SR$_{0.5}$, and SR$_{0.75}$, respectively.

TrackingNet (Muller et al. 2018). TrackingNet consists of over 30,000 video sequences collected from real-world scenarios, encompassing diverse object categories, motion patterns, and environmental conditions. It captures natural scene dynamics with varying camera movements and includes challenging cases such as fast-moving small objects, occlusions, and abrupt trajectory changes, ensuring a realistic evaluation setting. Tab.2 presents the results of MoDTrack along with several state-of-the-art trackers on the TrackingNet benchmark. MoDTrack$_{384}$ achieves an AUC score of 85.9%, showcasing its ability to track objects under diverse and dynamic conditions.

LaSOT (Fan et al. 2019). LaSOT is a large-scale long-term tracking dataset with 1,400 high-resolution videos averaging over 2,500 frames. It covers diverse objects and

Method	Source	LaSOT			LaSOT$_{\text{ext}}$			TrackingNet			TNL2K		UAV123
		AUC	P$_{norm}$	P	AUC	P$_{norm}$	P	AUC	P$_{norm}$	P	AUC	P	AUC
MoDTrack$_{256}$	Ours	**72.9**	**82.6**	**79.1**	**52.4**	62.9	**60.5**	**85.0**	**89.6**	84.4	**59.7**	**62.3**	70.5
TemTrack$_{256}$(Xie et al. 2024a)	AAAI25	72.0	82.1	**79.1**	**52.4**	**63.3**	60.2	84.3	88.8	83.5	58.8	60.9	**70.8**
LMTrack$_{256}$(Xu et al. 2025)	AAAI25	69.8	79.2	76.3	49.0	59.6	55.8	84.2	89.0	82.8	-	-	-
ARTrackV2$_{256}$(Bai et al. 2024)	CVPR24	71.6	80.2	77.2	50.8	61.9	57.7	84.9	89.3	**84.5**	59.2	-	69.9
AQATrack$_{256}$(Xie et al. 2024b)	CVPR24	71.4	81.9	78.6	51.2	62.2	58.9	83.8	88.6	83.1	57.8	59.4	70.7
EVPTrack$_{224}$(Shi et al. 2024)	AAAI24	70.4	80.9	77.2	48.7	59.5	55.1	83.5	88.3	-	57.5	58.8	70.2
ROMTrack$_{256}$(Cai et al. 2023)	ICCV23	69.3	78.8	75.6	48.9	59.3	55.0	83.6	88.4	82.7	56.9	58.1	69.7
SeqTrack-B$_{256}$(Chen et al. 2023)	CVPR23	69.9	79.7	76.3	49.5	60.8	56.3	83.3	88.3	82.2	54.9	-	69.2
MixFormer-22k$_{320}$(Cui et al. 2022)	CVPR22	69.2	78.7	74.7	-	-	-	83.1	88.1	81.6	-	-	70.4
OSTrack$_{256}$(Ye et al. 2022)	ECCV22	69.1	78.7	75.2	47.4	57.3	53.3	83.1	87.8	82.0	54.3	-	68.3
STARK-ST101$_{320}$(Yan et al. 2021a)	ICCV21	67.1	77.0	-	-	-	-	82.0	86.9	-	-	-	68.2
TransT$_{256}$(Chen et al. 2021)	CVPR21	64.9	73.8	69.0	-	-	-	81.4	86.7	80.3	50.7	51.7	69.1
Ocean$_{255}$(Zhang et al. 2020)	ECCV20	56.0	65.1	56.6	-	-	-	-	-	-	38.4	37.7	-
SiamRPN++$_{255}$(Li et al. 2019)	CVPR19	49.6	56.9	49.1	34.0	41.6	39.6	73.3	80.0	69.4	41.3	41.2	61.0
ECO$_{224}$ (Danelljan et al. 2017)	ICCV17	32.4	33.8	30.1	22.0	25.2	24.0	-	-	-	32.6	31.7	53.5
SiamFC$_{255}$ (Bertinetto et al. 2016)	ECCVW16	33.6	42.0	33.9	23.0	31.1	26.9	-	-	-	29.5	28.6	46.8
Some Trackers with Higher Resolution													
OSTrack$_{384}$(Ye et al. 2022)	ECCV22	71.1	81.1	77.6	50.5	61.3	57.6	83.9	88.5	83.2	55.9	56.3	70.7
ROMTrack$_{384}$(Cai et al. 2023)	ICCV23	71.4	81.4	78.2	51.3	62.4	58.6	84.1	89.0	83.7	58.0	59.6	70.5
F-BDMTrack$_{384}$(Yang et al. 2023)	ICCV23	72.0	81.5	77.7	50.8	61.3	57.8	84.5	89.0	84.0	57.8	59.4	70.9
SeqTrack-B$_{384}$(Chen et al. 2023)	CVPR23	71.5	81.1	77.8	50.5	61.6	57.5	83.9	88.8	83.6	56.4	-	68.6
ARTrack$_{384}$(Wei et al. 2023)	CVPR23	72.6	81.7	79.1	51.9	62.0	58.5	85.1	89.1	84.8	59.8	-	70.5
ODTrack-B$_{384}$(Zheng et al. 2024)	AAAI24	73.2	83.2	80.6	52.4	63.9	60.1	85.1	90.1	84.9	60.9	64.5	-
AQATrack$_{384}$(Xie et al. 2024b)	CVPR24	72.7	82.9	80.2	52.7	64.2	60.8	84.8	89.3	84.3	59.3	62.3	71.2
HIPTrack$_{384}$(Cai, Liu, and Wang 2024)	CVPR24	72.7	82.9	79.5	53.0	64.3	60.6	84.5	89.1	83.8	-	-	70.5
ARTrackV2-B$_{384}$(Bai et al. 2024)	CVPR24	73.0	82.0	79.6	52.9	63.4	59.1	85.7	89.8	85.5	-	-	-
TemTrack$_{384}$(Xie et al. 2024a)	AAAI25	73.1	83.0	80.7	**53.4**	**64.8**	61.0	85.0	89.3	84.8	-	-	-
MoDTrack$_{384}$	Ours	**74.4**	**83.8**	**81.5**	53.2	63.9	**61.3**	**85.9**	**90.2**	**86.0**	**61.8**	**66.3**	**71.8**

Table 2: Performance comparisons with state-of-the-art trackers on the test set of LaSOT, LaSOT$_{\text{ext}}$, TrackingNet, TNL2K, and UAV123. The top two results are highlighted using **bold** and underlined fonts respectively.

	SiamFC	OSTrack	SeqTrack	ARTrack	ROMTrack	F-BDMTrack	AQATrack	TemTrack	MambaLCT	MoDTrack$_{256}$	MoDTrack$_{384}$
AO	34.8	73.7	74.5	75.5	74.2	75.4	76.0	76.1	76.2	74.0	**76.3**
SR$_{0.5}$	35.3	83.2	84.3	84.3	76.0	84.3	85.2	84.9	**86.7**	84.1	86.4
SR$_{0.75}$	9.8	70.8	74.3	74.3	72.4	72.9	74.9	74.4	74.3	71.9	**75.6**

Table 3: Performance comparisons with state-of-the-art trackers on the test set of GOT-10k.

challenges like severe occlusion, abrupt motion, and background clutter. Unlike short-term datasets, LaSOT tests trackers on long sequences, handling target disappearance and reappearance. As shown in Tab.2, MoDTrack$_{256}$ improves AUC by 0.9% over TemTrack, demonstrating superior long-term tracking. Fig.5 shows attribute-based results, highlighting robustness in fast motion, clutter, and partial occlusion, validating our approach's effectiveness.

LaSOT$_{\text{ext}}$ (Fan et al. 2021). LaSOT$_{\text{ext}}$ extends LaSOT with more sequences, increasing diversity and difficulty for long-term tracking. It maintains high-quality annotations and long durations while introducing more categories and challenging cases such as small fast-moving targets, occlusion, and appearance changes. MoDTrack256 achieves a 52.4% AUC on LaSOT$_{\text{ext}}$, matching TemTrack.

TNL2K (Wang et al. 2021), UAV123 (Benchmark 2016), and OTB2015 (Wu, Lim, and Yang 2013). We further evaluate our tracker on three additional benchmarks: TNL2K, UAV123, and OTB2015. TNL2K is a high-quality multimodal dataset with 700 challenging videos and natural language annotations. UAV123 and OTB2015 are smaller-scale benchmarks containing 123 and 100 videos, respectively. As shown in Tab.2, MoDTrack sets new state-of-the-art results on all three datasets, demonstrating its robustness across diverse tracking scenarios.

Ablation Study and Analysis

We use MoDTrack$_{256}$ without all our designed modules as the baseline model in our ablation study. The baseline model's test results are shown in Tab.4 (#1).

Impact of Geometric Priors. To explore the effect of geometric priors, we conduct an extensive ablation study by fusing geometric cues from different layers. As shown in Tab.4 (#2), incorporating geometric priors from the final layer improves performance compared to the baseline. Fusing geometric priors from both an intermediate and final layer (#3) further enhances overall performance. This clearly demonstrates that additional geometric information enables better foreground-background separation and effectively improves model performance. The progressive fusion of geometric priors also facilitates more effective use of visual knowledge, significantly enhancing spatial awareness.

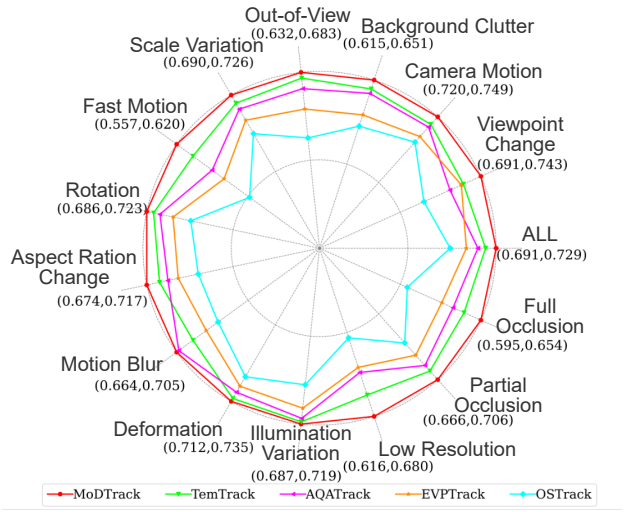

Figure 5: AUC scores of different attributes on LaSOT.

#	Method	AUC	△
1	Baseline	71.1	-
Impact of Depth Priors			
2	Fusion depth priors in $\{7\}$	71.3	+ 0.2
3	Fusion depth priors in $\{3, 7\}$	71.7	+ 0.6
Impact of Dictionary Size			
4	Dictionary Size = 100	72.2	+ 1.1
5	Dictionary Size = 200	72.5	+ 1.4
6	Dictionary Size = 400	72.6	+ 1.5
Impact of motion pattern Matching Methods			
7	Matching via Nearest Neighbor	72.6	+ 1.5
8	Matching via Attention Matrix	72.9	+ 1.8

Table 4: Ablation Study on LaSOT evaluating the impact of depth priors and Motion State Matching and Decoding Module, with △ indicating change from baseline.

Impact of Dictionary Size. We investigate how the dictionary size affects tracking performance. Tab.4 (#4) shows that by introducing the motion pattern dictionary, which represents different motion patterns as learnable features, the model's ability to mine and utilize motion information is significantly enhanced, resulting in a 1.1% performance improvement over the baseline. As shown in Tab.4 (#5 and #6), increasing the dictionary size further improves performance, as the greater diversity of motion patterns helps the model better adapt to various target movements. This analysis suggests that optimizing the diversity of motion patterns in the dictionary improves retrieval, leading to better matching accuracy in complex motion scenarios.

Impact of motion pattern Matching Methods. We compare motion pattern matching strategies to assess their impact on tracking performance. As detailed in Section 3.4, the attention matrix-based method (#8) selects the motion pattern with the highest similarity score, using a learnable mechanism to focus on relevant features. This process, enhanced by a projection layer that refines feature representations before matching, allows the model to better adapt to complex and evolving motion patterns. In contrast, the Nearest Neighbor Assignment method (#7) identifies matches by minimizing the squared Euclidean distance between features, offering a simple and efficient solution but lacking flexibility to cope with variations in feature distributions, and thus struggles in dynamic or cluttered scenarios. Overall, the attention-based strategy proves more effective and adaptive, delivering improved tracking accuracy by enabling robust retrieval of temporally varying motion cues.

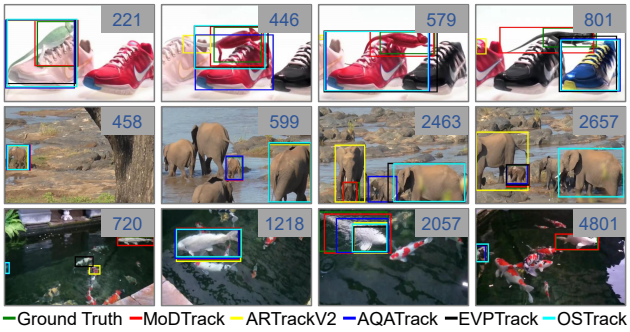

Figure 6: Qualitative comparison of our tracker with other trackers on LaSOT (Fan et al. 2019) benchmark.

Visualization and Qualitative Comparison. To demonstrate the effectiveness of our method in complex scenarios with environmental interference, such as similar object and background occlusions, we visualize tracking results of MoDTrack and four state-of-the-art trackers on the LaSOT dataset. As shown in Fig.6, this highlights the advantages of our approach. By incorporating geometric cues and motion pattern modeling, our method effectively reduces the impact of interference. Geometric cues help distinguish the target from the background, especially in occluded or cluttered scenes. Meanwhile, motion pattern modeling uses the target's historical motion to maintain accurate tracking despite rapid motion or similar object interference.

Conclusion

We present a tracker integrating motion pattern modeling and geometric priors to improve performance in dynamic, complex environments. By encoding motion patterns as learnable features, the model captures inter-frame motion relationships, enabling adaptation to rapid movements and complex object dynamics. The backbone extracts spatio-temporal features, while geometric priors enhance spatial awareness, facilitating foreground-background separation and robust occlusion handling. Extensive evaluations on seven benchmarks validate the method's effectiveness in reducing tracking drift and handling challenging conditions.

Acknowledgments

This work is supported by the Project of Guangxi Science and Technology (No.2025GXNSFAA069676, 2024GXNSFGA010001, and GuiKeFN2504240017), the National Natural Science Foundation of China (No.U23A20383, 62472109 and 62466051), Research Fund of Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education (EBME25-F-16), the Innovation Project of Guangxi Graduate Education (XYCS2025123), the Guangxi "Young Bagui Scholar" Teams for Innovation and Research Project, the Research Project of Guangxi Normal University (No. 2025DF001).

References

Bai, Y.; Zhao, Z.; Gong, Y.; and Wei, X. 2024. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19048–19057.

Benchmark, U. 2016. A benchmark and simulator for uav tracking. In *European conference on computer vision*, volume 7.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, 850–865. Springer.

Bill, J.; Gershman, S. J.; and Drugowitsch, J. 2022. Visual motion perception as online hierarchical inference. *Nature communications*, 13(1): 7403.

Cai, W.; Liu, Q.; and Wang, Y. 2024. HIPTrack: Visual Tracking with Historical Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19258–19267.

Cai, Y.; Liu, J.; Tang, J.; and Wu, G. 2023. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9589–9600.

Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022a. Backbone is all your need: A simplified architecture for visual object tracking. In *European Conference on Computer Vision*, 375–392. Springer.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14572–14581.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8126–8135.

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

Cheng, L.; Wang, J.; and Li, Y. 2022. ViTrack: Efficient Tracking on the Edge for Commodity Video Surveillance Systems. *IEEE Transactions on Parallel and Distributed Systems*, 33(3): 723–735.

Chetverikov, A.; and Jehee, J. F. 2023. Motion direction is represented as a bimodal probability distribution in the human visual cortex. *Nature Communications*, 14(1): 7634.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13608–13618.

Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; and Felsberg, M. 2017. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6638–6646.

Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Harshit; Huang, M.; Liu, J.; et al. 2021. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129: 439–461.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5374–5383.

Gao, S.; Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Learning dual-fused modality-aware representations for RGBD tracking. In *European Conference on Computer Vision*, 478–494. Springer.

Ge, J.; Cao, J.; Chen, X.; Zhu, X.; Liu, W.; Liu, C.; Wang, K.; and Liu, B. 2025. Beyond visual cues: Synchronously exploring target-centric semantics for vision-language tracking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5): 1–21.

Ge, J.; Cao, J.; Zhu, X.; Zhang, X.; Liu, C.; Wang, K.; and Liu, B. 2024. Consistencies are all you need for semi-supervised vision-language tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1895–1904.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025a. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3581–3589.

Hu, X.; Zhong, B.; Liang, Q.; Shi, L.; Mo, Z.; Tai, Y.; and Yang, J. 2025b. Adaptive Perception for Unified Visual Multimodal Object Tracking. *IEEE Transactions on Artificial Intelligence*, 6(10): 2819–2829.

Hu, X.; Zhong, B.; Liang, Q.; Zhang, S.; Li, N.; and Li, X. 2024a. Toward Modalities Correlation for RGB-T Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9102–9111.

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

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1562–1577.

Kart, U.; Kamarainen, J.-K.; and Matas, J. 2018. How to make an rgbd tracker? In *proceedings of the european conference on computer vision (ECCV) Workshops*, 0–0.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4282–4291.

Lin, T. 2017. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002*.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, 300–317.

Pereira, R.; Carvalho, G.; Garrote, L.; and Nunes, U. J. 2022. Sort and deep-SORT based multi-object tracking for mobile robotics: Evaluation with new data association metrics. *Applied Sciences*, 12(3): 1319.

Premachandra, C.; Ueda, S.; and Suzuki, Y. 2020. Detection and tracking of moving objects at road intersections using a 360-degree camera for driver assistance and automated driving. *IEEE Access*, 8: 135652–135660.

Qian, Y.; Yan, S.; Lukežič, A.; Kristan, M.; Kämäräinen, J.-K.; and Matas, J. 2021. DAL: A deep depth-aware long-term tracker. In *2020 25th International conference on pattern recognition (ICPR)*, 7825–7832. IEEE.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shi, L.; Zhong, B.; Liang, Q.; Li, N.; Zhang, S.; and Li, X. 2024. Explicit Visual Prompts for Visual Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4838–4846.

Song, Z.; Luo, R.; Yu, J.; Chen, Y.-P. P.; and Yang, W. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2321–2329.

Tian, Y.; Xie, L.; Qiu, J.; Jiao, J.; Wang, Y.; Tian, Q.; and Ye, Q. 2024. Fast-iTPN: Integrally pre-trained transformer pyramid network with token migration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, B.; Li, W.; and Ge, J. 2025. R1-Track: Direct Application of MLLMs to Visual Object Tracking via Reinforcement Learning. arXiv:2506.21980.

Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13763–13773.

Wang, Y.; Zhang, D.; Li, R.; Zheng, Z.; and Li, M. 2025. PD-SORT: Occlusion-Robust Multi-Object Tracking Using Pseudo-Depth Cues. *IEEE Transactions on Consumer Electronics*, 71(1): 165–177.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9697–9706.

Wu, Y.; Lim, J.; and Yang, M.-H. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2411–2418.

Xie, J.; Zhong, B.; Liang, Q.; Li, N.; Mo, Z.; and Song, S. 2024a. Robust Tracking via Mamba-based Context-aware Token Learning. *arXiv preprint arXiv:2412.13611*.

Xie, J.; Zhong, B.; Mo, Z.; Zhang, S.; Shi, L.; Song, S.; and Ji, R. 2024b. Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19300–19309.

Xu, C.; Zhong, B.; Liang, Q.; Zheng, Y.; Li, G.; and Song, S. 2025. Less is More: Token Context-aware Learning for Object Tracking. *arXiv preprint arXiv:2501.00758*.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021a. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10448–10457.

Yan, S.; Yang, J.; Käpylä, J.; Zheng, F.; Leonardis, A.; and Kämäräinen, J.-K. 2021b. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10725–10733.

Yang, D.; He, J.; Ma, Y.; Yu, Q.; and Zhang, T. 2023. Foreground-Background Distribution Modeling Transformer for Visual Object Tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10083–10093.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything V2. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 21875–21911. Curran Associates, Inc.

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 *European Conference on Computer Vision*, 341–357. Springer.

Zhang, D.; Fu, Y.; and Zheng, Z. 2022. UAST: Uncertainty-aware Siamese tracking. In *International Conference on Machine Learning*, 26161–26175. PMLR.

Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-aware anchor-free tracking. In *European conference on computer vision*, 771–787. Springer.

Zheng, Y.; Zhong, B.; Liang, Q.; Mo, Z.; Zhang, S.; and Li, X. 2024. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7588–7596.

Zheng, Y.; Zhong, B.; Liang, Q.; Zhang, S.; Li, G.; Li, X.; and Ji, R. 2025. Towards Universal Modal Tracking With Online Dense Temporal Token Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11): 10192–10209.