MambaLCT: Boosting Tracking via Long-term Context State Space Model

Xiaohai Li[1,2], Bineng Zhong[1*], Qihua Liang[1], Guorong Li[3], Zhiyi Mo[2†], Shuxiang Song[1]

[1]Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China

[2]Guangxi Colleges and Universities Key Laboratory of Intelligent Software, Wuzhou University, Wuzhou 543002, China

[3]Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing 100101, China

bruc_0619@stu.gxnu.edu.cn, bnzhong@gxnu.edu.cn, qhliang@gxnu.edu.cn

liguorong@ucas.edu.cn, zhiyim@gxuwz.edu.cn, songshuxiang@mailbox.gxnu.edu.cn

Abstract

Effectively constructing context information with long-term dependencies from video sequences is crucial for object tracking. However, the context length constructed by existing work is limited, only considering object information from adjacent frames or video clips, leading to insufficient utilization of contextual information. To address this issue, we propose MambaLCT, which constructs and utilizes target variation cues from the first frame to the current frame for robust tracking. First, a novel unidirectional Context Mamba module is designed to scan frame features along the temporal dimension, gathering target change cues throughout the entire sequence. Specifically, target-related information in frame features is compressed into a hidden state space through selective scanning mechanism. The target information across the entire video is continuously aggregated into target variation cues. Next, we inject the target change cues into the attention mechanism, providing temporal information for modeling the relationship between the template and search frames. The advantage of MambaLCT is its ability to continuously extend the length of the context, capturing complete target change cues, which enhances the stability and robustness of the tracker. Extensive experiments show that long-term context information enhances the model's ability to perceive targets in complex scenarios. MambaLCT achieves new SOTA performance on six benchmarks while maintaining real-time running speeds.

Code — https://github.com/GXNU-ZhongLab/MambaLCT

Introduction

Visual tracking involves locating the object in subsequent frames based on the given initial template information. In long and complex sequences, the target's appearance and motion state may change. However, traditional tracking algorithms (Chen et al. 2022; Hu et al. 2023; Ye et al. 2023) rely solely on the initial frame to locate the target in subsequent frames. They lack an understanding about change in objects through the sequence, leading to reduced stability and robustness of the tracker.

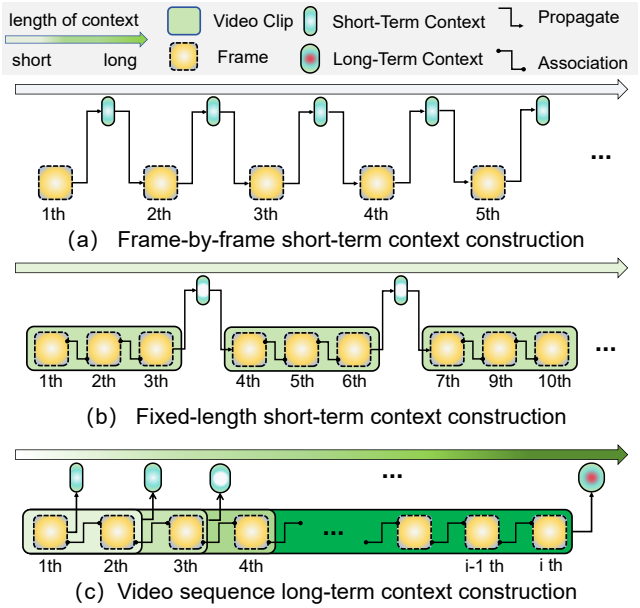

Figure 1: Comparison between current SOT context information construction paradigm and our method. (a) and (b) construct short-term context information within the scope of frames or video clips (Shi et al. 2024; Zheng et al. 2024). (c) Our propose MambaLCT analyzes the complete video sequence to construct long-term context information.

Recently, an increasing number of works have focused on the context information of a target in video sequences. Context information records the changes in the target's position, scale, appearance, and other trends. Effectively constructing context information from video sequences plays a crucial role in tracking algorithms. The length of the contextual information constructed in the current work is limited, which directly affects the range within which the tracker can perceive target changes. The current methods for constructing short-term context information can be summarized as follows: (1) *The frame-by-frame short-term context construction method*, as shown in Fig. 1(a). KeepTrack (Mayer et al. 2021) and EVPTrack(Shi et al. 2024) explores the associa-

*Corresponding author.

†Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tions between consecutive frames to enhance target perception. This association is a fragmented context, which focuses on target changes in adjacent frames. (2) *The fixed-length context construction method*, as shown in Fig. 1(b), models the temporal sequence of images within a fixed length. ODTrack (Zheng et al. 2024) and AQATrack (Xie et al. 2024) capture target context information using video clips and sliding windows, length of 4. The methods mentioned above rely on the Transformer for learning spatio-temporal information. While the Transformer performs exceptionally well in learning appearance features, its quadratic computational complexity restricts the length of context information it can handle. Recently, Mamba has shown great potential for computational efficiency in building long-term dependencies. Nevertheless, the performance of Mamba for non-autoregressive appearance feature learning is often underwhelming. Mamba and Transformer complement each other in learning appearance and context information, which has led us to consider: *Can our method combine the strengths of both to extend the context length, thereby enabling the learning of more robust appearance features?*

In this paper, we introduce a new tracking framework, namely MambaLCT, which aims to fully utilize target contextual information, as demonstrated in Fig. 1(c). MambaLCT gradually expands the range of context information to cover from the initial frame to the current frame. Specifically, we use Transformer in an autoregressive manner to learn the appearance features of the search image. Then, the autoregressive appearance features are continuously fed into the Context Mamba module during scanning frame features. Through Mamba's selective scanning mechanism, the information relevant to the target in the appearance features of historical search frames is continuously aggregated into the final state space via the transfer of hidden states. Compared to fragmented context information, unified modeling of all historical appearance features ensures the consistency and completeness of context information. Finally, the context information from the first frame to the current frame provides all historical target change information for modeling the appearance features of the next frame, thereby enhancing the accuracy and robustness of the tracker. The main contributions of this work are as follows:

- We propose a tracker named MambaLCT, which can effectively capture the long-term behavior and overall motion of the target, providing more comprehensive contextual information.
- We design a Context Mamba module that can effectively and with low resource consumption construct long-term contextual information along the temporal dimension.
- Our method has achieved a new state-of-the-art tracking performance on six visual tracking benchmarks, including LaSOT, LaSOT$_{\mathrm{ext}}$, GOT-10K, TrackingNet, TNL2K and UAV123.

Related Work

Tracking paradigm based on initial template. The initial template is reliable target information manually annotated. Whether it is the earlier two-stream framework trackers or the current single-stream framework trackers, they all heavily rely on the features of the initial template frame to complete tracking. Therefore, most tracking algorithms follow the search-template image matching paradigm. For example, SiamFC (Bertinetto et al. 2016) and SiamBAN (Chen et al. 2022), which are based on Siamese two-stream trackers, use networks with shared parameters to extract features from the search image and the template image separately. Then, they calculate the regions in the search image that are similar to the template image through cross-correlation operations. With the introduction of Transformers into the field of computer vision, an increasing number of works have started using Transformers for representation learning. An example of this is OSTrack (Ye et al. 2022), which addresses the issue of poor target perception caused by the two-stream, two-stage tracking framework. Although the tracking paradigm relying on initial template features has achieved competitive performance, with the advent of long sequence tracking benchmarks (Fan et al. 2019), the object features recorded by the initial template are insufficient to describe the target in subsequent frames. In this paper, our aim is to construct a tracker that can perceive long-term target changes.

Tracking paradigm based on context information. Video sequences contain rich contextual information, recording the state of the target at multiple points in time. To overcome the limitations of the initial template, tracking frameworks that heavily utilize contextual information have begun to emerge. Such as VideoTrack (Xie et al. 2023) proposed a method using a video Transformer to encode spatio-temporal information in videos. Some exciting works were proposed last year, utilizing frame-by-frame propagation methods to transmit cross-frame contextual information. ODTrack (Zheng et al. 2024) uses a token to transmit cross-frame information from the first frame to the last frame. The token learns cross-frame contextual information by being encoded together with image pairs. This method is simple and efficient, but it can lead to redundant computations. AQATrack (Xie et al. 2024) designs a Transformer-based spatio-temporal information fusion module to extract information between frames and propagate it through queries. The context length constructed by these aforementioned methods can only focus on target changes in adjacent frames or within a window, resulting in limited help from the context for the tracker. To address this issue, we propose a method for constructing context that can focus on target changes in all historical search frames.

Mamba for computer vision. Mamba model has demonstrated excellent semantic modeling capabilities in the field of natural language processing. Some works (Zhu et al. 2024; Huang et al. 2024) have successfully applied the Mamba model to the field of computer vision. Compared to the quadratic complexity of Transformers, Mamba has great potential in handling long sequences of data with its linear complexity. Vim (Zhu et al. 2024) designed a bidirectional state space to enhance Mamba's sensitivity to visual data and meet the contextual requirements of visual understanding. (Li et al. 2024) proposed VideoMamba, which introduces Mamba into the video domain, addressing issues of local redundancy and global dependencies in video understanding.

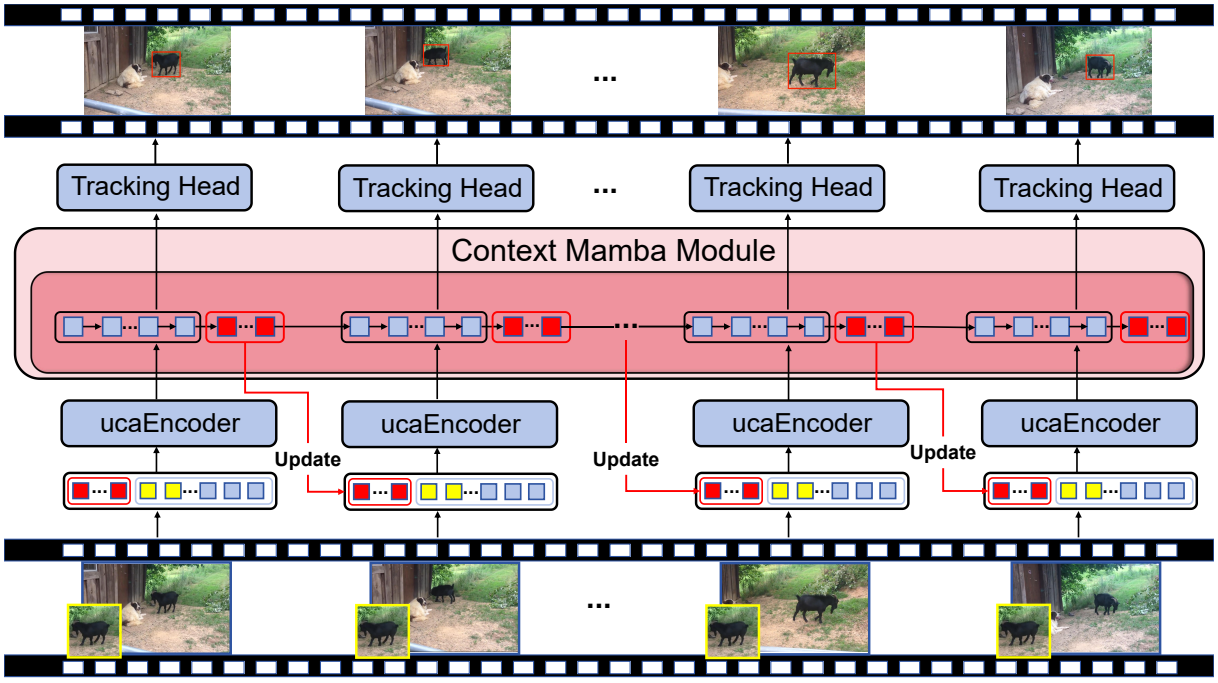

Figure 2: Overview of our framework. The input video frames are converted into tokens through patch embedding. Then, these tokens, along with the contextual information, are fed into the ucaEncoder for unified modeling of the contextual and appearance information. During the temporal scanning process, the representational information of the images is continuously fed into the Context Mamba module to construct the target's change cues.

Mamba4D (Liu et al. 2024) designed an efficient 4D point cloud video backbone framework to unify the irregular and unordered distribution of points. Due to the inherent mechanism of SSM, Mamba has autoregressive properties. However, appearance features in tracking tasks do not conform to this characteristic. In this work, we separate the modeling of temporal and spatial information in tracking tasks, applying Mamba to the temporal modeling.

Method

In this section, we introduce how MambaLCT constructs and utilizes long-term contextual. First, we briefly present our framework, followed by an explanation of how we construct contextual information along the temporal dimension and unify the modeling of appearance and context information. Finally, we describe the training pipeline.

Overview

An overview of the MambaLCT pipeline is shown in Fig. 2. This framework is very straightforward, consisting of three main components: the ucaEncoder, the Context Mamba module, and the tracking head. First, the initial template $t \in \mathbb{R}^{3\times H_t\times W_t}$ and the video frames $s \in \mathbb{R}^{3\times H_s\times W_s}$ are input into the pipeline sequentially, H and W represent the height and width of the image. s and t are converted into 1D tokens $s_p \in \mathbb{R}^{N_s\times D}$ and $t_p \in \mathbb{R}^{N_t\times D}$ through the image patch embedding process. Here $N_s = H_sW_s/16^2$, $N_t = H_tW_t/16^2$, $D = 512$. These tokens are then concatenated with the contextual information tokens $c_p \in \mathbb{R}^{N_c\times D}$ and fed into the ucaEncoder, as shown in Fig. 3, N_c is the length of c_p. Then, we design a Context Mamba Module to construct long-term contextual information. The search frame feature flow output by the ucaEncoder is continuously fed into the Context Mamba module, where the target information with long-term dependencies is aggregated into the c_p. Finally, the search image features enter the tracking head for classification and regression to obtain the results.

Preliminaries

Mamba excels at handling sequences with long-term dependencies. The Mamba model is inspired by State Space Models. SSMs are a class of models used for handling time series data, typically mapping an input sequence $x(t) \in \mathbb{R}$ an output sequence $y(t) \in \mathbb{R}$. The mapping process is as follows:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $h(t)$ is the hidden state, $\mathbf{A}$ is the evolution parameter, and $\mathbf{B}$, $\mathbf{C}$ are the projection parameters.

To effectively apply the Mamba model in the field of deep learning algorithms, we need to discretize the continuous parameters $\mathbf{A}$ and $\mathbf{B}$. A common discretization method is zero-order hold (ZOH), which is defined as follows:

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

where Δ is a timescale parameter for discretization. The dis-

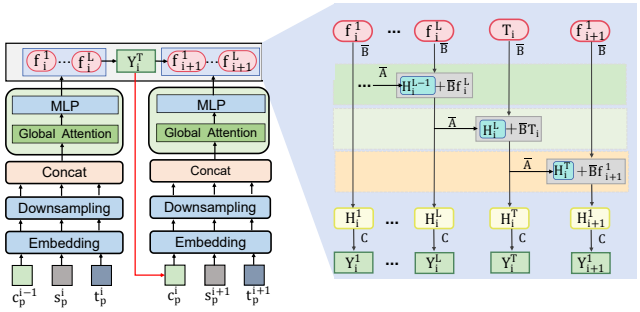

Figure 3: Illustration of the process of constructing and propagating context information. On the left is the structure of the ucaEncoder, and on the right is the process of constructing contextual information.

cretized version of Eq. (1) and Eq. (2) can be rewritten as:

$$\begin{aligned} h_t &= \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} x_t, \\ y_t &= \mathbf{C} h_t. \end{aligned} \tag{3}$$

Context Mamba Module

Existing works primarily extract contextual information from adjacent frames or images within a nearby window. (Zheng et al. 2024; Xie et al. 2024) proposing a novel method to construct context information. They increase the context length by focusing on target changes within a fixed window. Their approach to constructing contextual information can be summarized as follows:

$$T \leftarrow f : \{T_{i-l}, ...T_i, S_{i-l}, ..., S_i\}, \tag{4}$$

where the commonly used f is generally based on Transformer methods. S represents the search image, T denoted the template image. l is the length of context information.

To fully utilize the contextual information in videos, we propose a Mamba-based context construction method, which can be described as follows:

$$T \leftarrow Mamba : \{S_1, S_2, S_3..., S_i\}, \tag{5}$$

stick to Mamba's strengths, we do not use Mamba to learn the spatial features of objects. Instead, we directly perform sequential modeling on the historical search frame flow. From Eq. (5) and Eq. (4), we can see that short-term contextual information is limited to the target information of the adjacent l frames, whereas our method can capture target information from the first frame to the current frame.

As shown in Fig. 3, the search features from the first frame to the i frame are continuously fed into the Context Mamba Module, as follows:

$$input = [f_1^1 W; \cdots; f_1^L W; \cdots; f_i^1 W; \cdots; f_i^L W], \tag{6}$$

where f_i^j represents the j-th token in the feature of the i-th frame, W is the learnable linear projection matrix, L is the length of image token. Through the selective scanning mechanism, the information related to the target in the f_i is compressed into the H_i^L. This process is described as:

$$H_i^t = \overline{\mathbf{A}} H_i^{t-1} + \overline{\mathbf{B}} f_i^t, \quad t \leq L \tag{7}$$

where H_i^t represents the current aggregated hidden state. Information about the target from frames 1 to i in the video sequence is also continuously aggregated into Y_i^T through the hidden state, Context information is transmitted from f_i to f_{i+1} through Y_i^T, as follows:

$$\begin{aligned} H_i^T &= \overline{\mathbf{A}} H_i^L + \overline{\mathbf{B}} T_i, \\ H_{i+1}^1 &= \overline{\mathbf{A}} H_i^T + \overline{\mathbf{B}} f_{i+1}^1, \\ Y_i^T &= \mathbf{C} H_i^T, \end{aligned} \tag{8}$$

where T_i is an empty token used to record all historical target information from the past i frames, and H_i^T serves as the medium for information transmission between frames. Y_i^T is used to update the c_p.

Unified Context and Appearance Modeling

The ucaEncoder is a critical component of our framework, capable of achieving unified modeling of contextual and appearance information. The structure of this encoder is shown in Fig. 3, and it is implemented based on a transformer architecture. Unlike the vanilla ViT, which directly downsamples the input image with a stride of 16, we apply a hierarchical ViT to perform multiple stages of downsampling to avoid potential information loss caused by such a large stride. Specifically, we inject c_p into the attention operations of S_p and t_p. Before this, we first perform two stages of downsampling on these three inputs, followed by global attention operations. This process can be summarized as follows:

$$\begin{aligned} i_c, i_s, i_t &= Ds(c_p, s_p, t_p), \\ f_{input} &= Concat(i_c, i_s, i_t), \\ f_i &= MLP(attention(f_{input})), \end{aligned} \tag{9}$$

where Ds denotes the downsampling operation, f_i represents the features of the current frame. Y_i^T, which contains all target change information from the first i frames, introduces contextual information into the relationship modeling between the template and search frames by updating c_p. This effectively enhances the target change cues, thereby improving the tracker's perception of the target.

Training and Loss Function

Common tracking dataset sampling methods involve randomly selecting video frames from a random dataset. These random images may come from different video sequences, making it impossible to learn context information. Our sampling method involves selecting video clips of a certain length from sequences within a random dataset, ensuring that there is contextual connection between the images. The context tokens from the previous frame, search frame patches, and template frame patches are concatenated and input into the ucaEncoder for interaction. Finally, the image features are fed into the tracking head for bounding box prediction. The tracking head consists of classification and regression head. We use classification loss, regression loss, and GIoU loss as training constraints. The total loss can be formulated as:

$$L = L_{cls} + \lambda_1 L_1 + \lambda_2 L_{GIoU}, \tag{10}$$

where λ_1 and λ_2 are two manually set loss weights.

Experiments

Implementation Details

Experimental environment. Our tracker implementation is based on Python 3.8 and Pytorch 1.13.1. Training and testing were conducted on two NVIDIA A100 GPUs. The tracking speed test was performed on a Tesla V100.

Tracker	Type	Resolution	Params	MACs	Speed	Device
SeqTrack-384B	ViT-B	384×384	89M	148G	12.8fps	Tesla V100
AQATrack-256	HiViT-B	384×384	72M	25.8G	67.6fps	Tesla V100
AQATrack-384	HiViT-B	384×384	72M	58.3G	44.2fps	Tesla V100
Our-256	HiViT-B	256×256	72M	25G	58.6fps	Tesla V100
Our-384	HiViT-B	384×384	72M	58G	45.3fps	Tesla V100

Table 1: Comparison of model parameters, MACs(G), and Spees(fps). These test results were obtained on the same machine.

Model parameter. To demonstrate the robustness of our tracker to different image scales, we designed two different trackers with varying image input resolutions, as shown in Tab. 1.

Training Details. During the training process, we set the video clip sampling length to 2 and the sampling quantity to 30000. The datasets used for training are GOT10K (Huang, Zhao, and Huang 2019), LaSOT (Fan et al. 2019), COCO (Lin et al. 2014), and TrackingNet (Muller et al. 2018). We use HiViT-Base (Zhang et al. 2023) as the backbone network to extract visual features, with its parameters initialized using MAE (He et al. 2022). The length of the cross-frame token is set to 1. The Mamba cross-frame information learning network uses Vim-Small (Zhu et al. 2024) and is initialized with its checkpoint. We made some modifications to Vim, employing a unidirectional scanning method. During training, MambaLCT uses the AdamW (Loshchilov and Hutter 2017) optimizer to adjust model parameters. The learning rate for the backbone network is set to 2×10^{-4}, while other parameters have a learning rate ten times higher. We train the model for a total of 300 epochs, and learning rate decay begins at the 240th epoch, with a decay rate set to 1×10^{-4}. The same learning rate and decay rate are applied when training on the GOT10K dataset, but the model is trained for only 150 epochs, with the decay starting at the 120th epoch. The batch size for training is set to 16. In the loss function, $\lambda_1 = 5$ and $\lambda_2 = 2$.

Inference Details. Unlike most tracking methods, in our frame-by-frame testing process, we use the search frame features and cross-frame tokens to construct a cross-frame information flow that includes the entire test sequence's image information. The cross-frame tokens act as a medium for information transmission between frames. Thanks to the Mamba model's advantages in handling long sequence data, our tracker achieves excellent speed and performance, as shown in Tab. 1.

State-of-the-art Comparisons

LaSOT. The LaSOT dataset contains 1400 high-quality video sequences with a total duration of over 3800 minutes. Each video sequence has an average length of approximately 2500 frames. The videos in this dataset encompass rich cross-frame information. With similar resource consumption, our model achieved new state-of-the-art performance at different image input resolutions. As shown in Tab. 1, MambaLCT-256 achieved 71.8%, 83.0%, and 79.4% in AUC, P_{Norm}, and precision score, respectively, across the three evaluation metrics. Compared to ODTrack-B (Zheng et al. 2024), MambaLCT-384 achieved improvements of 0.4%, 0.9%, and 1.0% in AUC, P_{Norm}, and precision score, respectively.

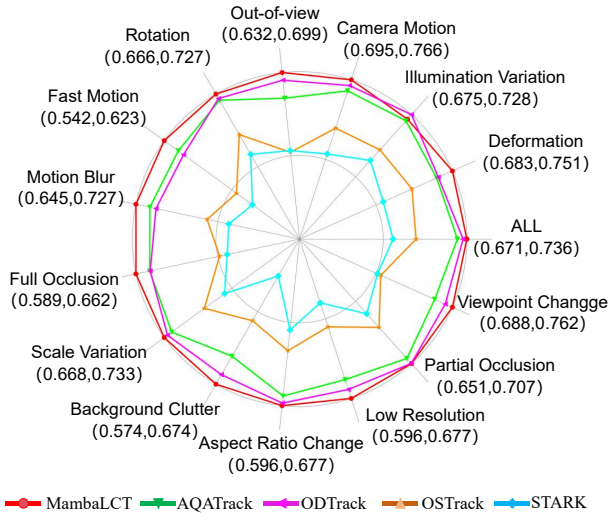

Figure 4: Attribute-based evaluation on the LaSOT test set. AUC score is used to rank different trackers.

LaSOT$_{ext}$. LaSOT$_{ext}$ is an extension supplement to the LaSOT dataset, encompassing 15 categories and 150 videos. Our proposed method for building long-term dependency cross-frame contextual information performs excellently in long video sequences. Our MambaLCT model achieved state-of-the-art performance with an input image resolution of 256, attaining 51.6% AUC, 64.0% P_{Norm}, and 59% precision score. MambaLCT-384 achieved an improvement of 0.6% in all three evaluation metrics compared to AQATrack-384. Compared to trackers using the same backbone network and input resolution, this result demonstrates that our proposed method offers superior performance.

GOT-10K. The GOT-10k contains 10,000 high-quality video sequences, covering 580 object categories. The dataset includes over 1.5 million bounding box annotations in total. The dataset is evaluated using two metrics: Average Overlap (AO) and Success Rate (SR). As shown in Tab. 2, our tracker demonstrated competitive performance across multiple metrics. MambaLCT-256 achieved 74.8%, 85.4%, and 72.1% in metrics AO, $SR_{0.5}$, and $SR_{0.75}$, respectively. The high-resolution MambaLCT surpassed ARTrack-384, showing increases of 1.7%, 2.4%, and 2.9% in terms AO, $SR_{0.5}$, and $SR_{0.75}$, respectively.

TrackingNet. The TrackingNet comprises over 30,000 video sequences, totaling more than 14 hours of video content. Each video sequence has an average length of 500 frames. The dataset particularly focuses on the stability and

Method	Source	LaSOT			$\text{LaSOT}_{\text{ext}}$			GOT-10K*			TrackingNet		
		AUC	P_{norm}	P	AUC	P_{norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	AUC	P_{norm}	P
MambaLCT-256	our	**71.8**	**83.0**	**79.4**	**51.6**	**64.0**	**59.0**	**74.8**	**85.4**	**72.1**	**84.3**	**89.2**	**83.9**
AQATrack-256(Xie et al. 2024)	CVPR24	71.4	81.9	78.6	51.2	62.2	58.9	73.8	83.2	72.1	83.8	88.6	83.1
F-BDMTrack-256(Yang et al. 2023)	ICCV23	69.9	79.4	75.8	47.9	57.9	54.0	72.7	82.0	69.9	83.7	88.3	82.6
ROMTrack-256(Cai et al. 2023)	ICCV23	69.3	78.8	75.6	48.9	59.3	55.0	72.9	82.9	70.2	83.6	88.4	82.7
ARTrack-256(Wei et al. 2023)	CVPR23	70.4	79.5	76.6	46.4	56.5	52.3	73.5	82.2	70.9	84.2	88.7	83.5
GRM(Gao, Zhou, and Zhang 2023)	CVPR2023	69.9	79.3	75.8	-	-	-	73.4	82.9	70.4	84.0	88.7	83.3
OSTrack-256(Ye et al. 2022)	ECCV22	69.1	78.7	75.2	47.4	57.3	53.3	71.0	80.4	68.2	83.1	87.8	82.0
VideoTrack(Xie et al. 2023)	CVPR23	70.2	-	76.4	-	-	-	72.9	81.9	69.8	83.8	88.7	83.1
AiATrack-320(Gao et al. 2022)	ECCV22	69.0	79.4	73.8	-	-	-	69.6	80.0	63.2	82.7	87.8	80.4
MixFormer-22k(Cui et al. 2022)	CVPR22	69.2	78.7	74.7	-	-	-	70.7	80.0	67.8	83.1	88.1	81.6
STARK(Yan et al. 2021)	ICCV21	67.1	77.0	-	-	-	-	68.8	78.1	64.1	81.3	86.1	-
TransT (Chen et al. 2021)	CVPR21	64.9	73.8	69.0	-	-	-	67.1	76.8	60.9	81.4	86.7	80.3
Ocean (Zhang et al. 2020)	ECCV 20	56.0	65.1	56.6	-	-	-	61.1	72.1	47.3	-	-	-
SiamRPN++(Li et al. 2019)	CVPR19	49.6	56.9	49.1	34.0	41.6	39.6	51.7	61.6	32.5	73.3	80.0	69.4
ECO (Danelljan et al. 2017)	ICCV 17	32.4	33.8	30.1	22.0	25.2	24.0	31.6	30.9	11.1	55.4	61.8	49.2
SiamFC (Bertinetto et al. 2016)	ECCVW16	33.6	42.0	33.9	23.0	31.1	26.9	34.8	35.3	9.8	57.1	66.3	53.3
Some Trackers with Higher Resolution													
OSTrack-384(Ye et al. 2022)	ECCV22	71.1	81.1	77.6	50.5	61.3	57.6	73.7	83.2	70.8	83.9	88.5	83.2
SeqTrack-B384(Chen et al. 2023)	CVPR23	71.5	81.1	77.8	50.5	61.6	57.5	74.5	84.3	71.4	83.9	88.8	83.6
ROMTrack-384(Cai et al. 2023)	ICCV23	71.4	81.4	78.2	51.3	62.4	58.6	74.2	84.3	72.4	84.1	89.0	83.7
F-BDMTrack-384(Yang et al. 2023)	ICCV23	72.0	81.5	77.7	50.8	61.3	57.8	75.4	84.3	72.9	84.5	89.0	84.0
ARTrack-384(Wei et al. 2023)	CVPR23	72.6	81.7	79.1	51.9	62.0	58.5	75.5	84.3	74.3	85.1	89.1	84.8
ODTrack(Zheng et al. 2024)	AAAI24	73.2	83.2	80.6	52.4	63.9	60.1	**77.0**	**87.9**	**75.1**	85.1	**90.1**	84.9
AQATrack-384(Xie et al. 2024)	CVPR24	72.7	82.9	80.2	52.7	64.2	60.8	76.0	85.2	74.9	84.8	89.3	84.3
MambaLCT-384	Ours	**73.6**	**84.1**	**81.6**	**53.3**	**64.8**	**61.4**	76.2	86.7	74.3	**85.2**	89.8	**85.2**

Table 2: Comparison with state-of-the-arts on four popular benchmarks: LaSOT, $\text{LaSOT}_{\text{ext}}$, GOT-10K, and TrackingNet. Where * denotes for trackers only trained on GOT10K. Best in bold, second best underlined.

persistence of targets in long video sequences. As depicted in Tab. 2, Our tracker achieved an AUC of 85.2%, a P_{Norm} of 89.8%, and a precision score of 83.9%. The performance on the TrackingNet dataset clearly demonstrates the superiority of our tracker in long-term tracking.

TNL2K. TNL2K is a large-scale dataset for natural language tracking, containing approximately 2,000 video sequences, with a training and testing split ratio of 13:7. One notable feature of TNL2K is that it consists of multi-source data, including RGB-T, cartoons, and more. As illustrated in Tab. 3, MambaLCT exhibited competitive performance, achieving an AUC score of 58.5%.

UAV123. UAV123 is a dataset for low-altitude UAV object tracking, comprising a total of 123 video sequences. As evidenced by the Tab. 3, compared to SeqTrack and OStrack, MambaLCT achieved improvements of 1.5% and 1.8% in AUC scores, respectively.

Ablation Study

To explore the effectiveness of our proposed method, we conducted extensive experiments using the MambaLCT-256 model on the LaSOT dataset.

Baseline Vs MambaLCT. Our MambaLCT is based on the OSTrack framework. Specifically, we replaced ViT with HiViT, used sequence sampling instead of single-frame sampling, and enhanced object perception within the background by learning video context information through Mamba. In Tab. 4(a), #1 indicates that the hierarchical ViT significantly improves our model's ability to extract spatial information, with a 0.6% increase in the AUC score. According to the results of #2 and #3, we can observe that using video sampling alone does not improve the performance of the tracker. However, as shown in #4, effectively utilizing context information from video sampling can significantly enhance the performance of the tracker, achieved 71.8%, 83.0%, and 79.4% in AUC, P_{Norm}, and precision score.

Study on length of $\boldsymbol{c}_p$. To verify the impact of context information token length on experimental performance, we designed a series of experiments using different lengths of $\boldsymbol{c}_p$. As shown in Tab. 4(b), when the length of CP is 1, the tracker achieved 71.8%, 83.0%, and 79.4% on the three metrics, respectively. With the length increases, the performance gradually decreases. We believe that this phenomenon is due to the autoregressive nature of Mamba, where each token's generation depends on the preceding tokens. $\boldsymbol{c}_p$ composed of too many consecutive tokens can lead to redundant contextual information.

Study on different Mamba layers. We evaluated the different insertion layers of the Mamba module we used and summarized the results in Tab. 4(c). Compared to inserting Mamba into the first three and last three layers of the backbone network, the method of uniformly inserting Mamba throughout the network can learn multi-level contextual information, achieving a maximum AUC score of 71.8%.

	SiamFC	MDNet	Ocean	TransT	TransInMo	ATOM	DiMP50	STARK	JointNLT	OSTrack	SeqTrack	Ours
TNL2K	29.5	38.0	38.4	50.7	52.0	44.7	44.7	-	56.9	55.9	56.4	**58.5**
UAV123	46.8	52.8	57.4	68.1	71.1	64.3	64.3	68.2	-	68.3	68.6	**70.1**

Table 3: Comparison with state-of-the-art methods on TNL2K and UAV123 benchmarks in AUC score.

(a) Study on our method					(b) Study on c_p length				(c) Study on Mamba Layers			
Num	Method	AUC	P_{norm}	P	Sampling Length	AUC	P_{norm}	P	Mamba Layers	AUC	P_{norm}	P
1	baseline	69.1%	78.7%	75.2%	1	**71.8%**	**83.0%**	**79.4%**	None	70.1%	80.2%	78.6%
2	+HiViT	70.5%	80.7%	77.5%	2	71.3%	82.7%	79.1%	(1,2,3)	70.5%	81.9%	77.5%
3	+sequen sampling	70.1%	80.2%	78.6%	3	70.8%	82.1%	78.0%	(3,6,9)	**71.8%**	**83.0%**	**79.4%**
4	+Mamba	**71.8%**	**83.0%**	**79.4%**	4	69.8%	81.0%	76.7%	(18,19,20)	70.9%	82.1%	78.0%

Table 4: Three ablation studies on the LaSOT benchmark.

Figure 5: On the LaSOT benchmark, we visualized the comparison results of our tracker with three SOTA trackers across three challenges.

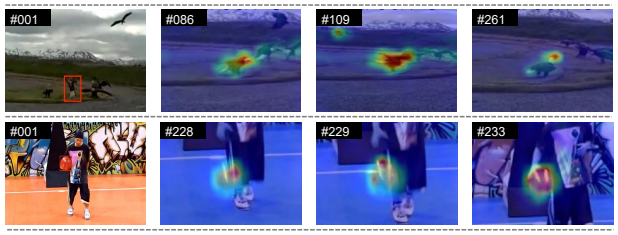

Figure 6: Visualization of the attention to the search frame by the cross-frame context information. As the length of the context information increases, the attention to the target becomes more focused.

Visualization and Limitation

Visualization. To visually demonstrate the effectiveness of our proposed method, we visualized the AUC score of MambaLCT under different challenges on the LaSOT benchmark, as shown in Fig. 4. Compared to other video-level trackers, MambaLCT excels in challenges such as motion blur, full occlusion, and deformation. As illustrated in Fig. 5, we visualized the tracking results of sequences under three challenges. From these results, we can see that thanks to the increased length of context information, our tracker can more accurately locate the target in complex situations such as target deformation and occlusion, compared to other trackers.

Additionally, to verify whether the context information can capture the target's information, we visualized the attention maps of the context information on the search frame, as shown in Fig. 6. We found that as the sequence length increases, the attention of the context information becomes more focused on the target, which also demonstrates the advantages of long-term context information.

Limitation. Our proposed video-level context information modeling method effectively captures the complete behavior and contextual understanding of the target, thereby enhancing the model's perception of the target. Despite achieving significant results, our proposed method faces limitations due to computational resource constraints. Consequently, the training and testing phases cannot be unified. During the training phase, we do not model the entire video sequence but instead sample portions of the sequence. Designing a more coherent training strategy is a direction for our future work. One feasible direction is to replace the previous image sampling with sequence sampling.

Conclusion

In this paper, we introduced MambaLCT, a novel tracking framework designed to enhance object perception by constructing long-term context information. We extended the context information length from the initial frame to the current frame, which provides a more comprehensive overview of the target's historical information. We combine the strengths of Mamba and Transformers to construct and utilize long-term contextual information. The Transformer learns the appearance features of the images in an autoregressive manner, while Mamba extracts long-term contextual information about the target from historical appearance data. Detailed experiments show that our MambaLCT achieves excellent results across six tracking benchmarks, although our work still has limitations. We hope this work can contribute valuable insights and assistance to current tracking paradigms.

Acknowledgments

This work is supported by the Project of Guangxi Science and Technology (No.2024GXNSFGA010001 and 2022GXNSFDA035079), the National Natural Science Foundation of China (No.U23A20383, 62472109 and 62466051), the Guangxi "Young Bagui Scholar" Teams for Innovation and Research Project, the Research Project of Guangxi Normal University (No.2024DF001), the Innovation Project of Guangxi Graduate Education (YCBZ2024083), and the grant from Guangxi Colleges and Universities Key Laboratory of Intelligent Software (No.2024B01).

References

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.

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, 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. 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 *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.; 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.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, 146–164. Springer.

Gao, S.; Zhou, C.; and Zhang, J. 2023. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18686–18695.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, X.; Zhong, B.; Liang, Q.; Zhang, S.; Li, N.; Li, X.; and Ji, R. 2023. Transformer Tracking via Frequency Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*.

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.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*.

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.

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 preprint arXiv:2403.06977*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, J.; Han, J.; Liu, L.; Aviles-Rivero, A. I.; Jiang, C.; Liu, Z.; and Wang, H. 2024. MAMBA4D: Efficient Long-Sequence Point Cloud Video Understanding with Disentangled Spatial-Temporal State Space Models. *arXiv preprint arXiv:2405.14338*.

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

Mayer, C.; Danelljan, M.; Paudel, D. P.; and Van Gool, L. 2021. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13444–13454.

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.

Shi, L.; Zhong, B.; Liang, Q.; Li, N.; Zhang, S.; and Li, X. 2024. Explicit Visual Prompts for Visual Object Tracking. *arXiv preprint arXiv:2401.03142*.

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.

Xie, F.; Chu, L.; Li, J.; Lu, Y.; and Ma, C. 2023. Videotrack: Learning to track objects via video transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22826–22835.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19300–19309.

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

Yang, D.; He, J.; Ma, Y.; Yu, Q.; and Zhang, T. 2023. Foreground-background distribution modeling transformer for visual object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10117–10127.

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.

Ye, J.; Zhong, B.; Liang, Q.; Zhang, S.; Li, X.; and Ji, R. 2023. Positive-Sample-Free Object Tracking via a Soft Constraint. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, X.; Tian, Y.; Xie, L.; Huang, W.; Dai, Q.; Ye, Q.; and Tian, Q. 2023. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations*.

Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-aware anchor-free tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 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. *arXiv preprint arXiv:2401.01686*.

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 preprint arXiv:2401.09417*.