Depth Attention for Robust RGB Tracking

Yu Liu[1][0009−0009−5898−0113], Arif Mahmood[2][0000−0001−5986−9876], and
Muhammad Haris Khan[3][0000−0001−9746−276X]

[1] Xinjiang University
7501847851y@gmail.com
[2] Information Technology University
arif.mahmood@itu.edu.pk
[3] Mohamed bin Zayed University of Artificial Intelligence
muhammad.haris@mbzuai.ac.ae

Abstract. RGB video object tracking is a fundamental task in computer vision. Its effectiveness can be improved using depth information, particularly for handling motion-blurred target. However, depth information is often missing in commonly used tracking benchmarks. In this work, we propose a new framework that leverages monocular depth estimation to counter the challenges of tracking targets that are out of view or affected by motion blur in RGB video sequences. Specifically, our work introduces following contributions. To the best of our knowledge, we are the first to propose a depth attention mechanism and to formulate a simple framework that allows seamlessly integration of depth information with state of the art tracking algorithms, without RGB-D cameras, elevating accuracy and robustness. We provide extensive experiments on six challenging tracking benchmarks. Our results demonstrate that our approach provides consistent gains over several strong baselines and achieves new SOTA performance. We believe that our method will open up new possibilities for more sophisticated VOT solutions in real-world scenarios. Our code and models are publicly released: https://github.com/LiuYuML/Depth-Attention.

Keywords: Visual Object Tracking · Multi-Modal Tracking · Monocular Depth Estimation · Single object tracking.

1 Introduction

RGB video object tracking[5,27,36,9] is a fundamental task in computer vision, aimed at accurately tracking a target object across video frames. Despite advancements in score map generation[5,27,1] and scale adaptation techniques [16,2,15,68,36], current SOTA algorithms[49,53,42,23,22,55] face challenges in scenarios with invisible target, motion blur, and fast motion.

These challenges may be addressed by employing depth information in the tracking algorithms. This integration makes the sudden object disappearances detectable through depth variation, while motion blur and fast motion have minimal impact on the depth pattern. However, incorporating RGB-D input for
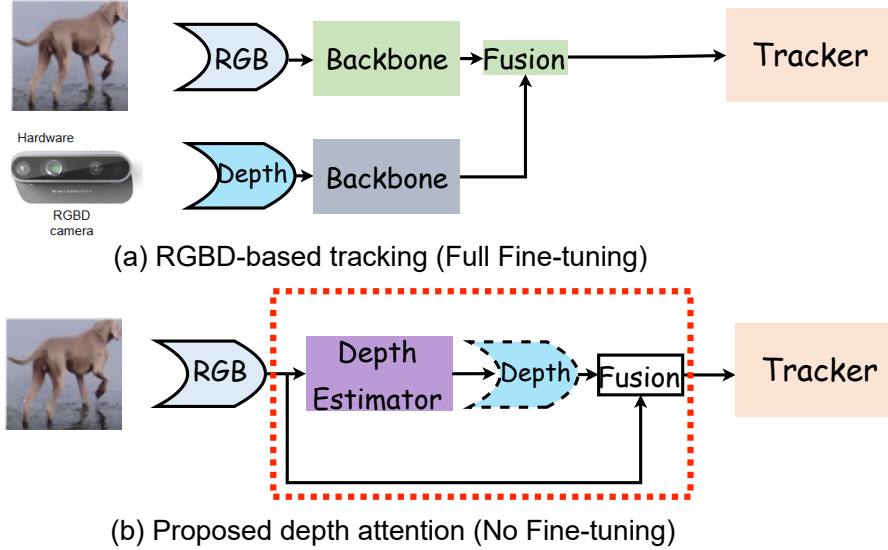
Fig. 1. Comparison of our approach (depth attention) with RGB-D Tracking.

video tracking is usually costly and falls into the category of RGB-D video object tracking, as highlighted in various studies[50,63,60,62]. On the other hand, the present RGB datasets employed for RGB video object tracking lacks depth information. To address this limitation and leverage the access to depth information in RGB Tracking, this paper introduces a simple tracking framework at the core of which is a new depth attention module specifically designed for RGB tracking algorithms. We would like to emphasize that the proposed module aims to enhance the performance of RGB tracking algorithms on the RGB datasets. Since the estimated monocular depth information is inherently less precise than the depth information from RGB-D sensors, we choose not to modify the architecture of existing RGB tracking methods as it is in the existing RGB-D algorithms [31,32,40,67]. Instead, we present a simple framework that can be seamlessly integrated into existing RGB tracking algorithms, resulting in improved performance on RGB datasets (see Figure 1). We summarize our key contributions as follows:

– To the best of our knowledge, we are the first to leverage depth information for improving RGB Tracking in a principled manner.
– Our approach is neither dependent on RGB-D datasets nor requires precise depth information from the RGB-D sensors. Our proposed depth attention efficiently leverages rapid monocular depth estimation and can be seamlessly incorporated into existing RGB Tracking algorithms.
– The effectiveness of our proposed method is validated on six challenging benchmarks. The results consistently demonstrate the state-of-the-art (SOTA) performance across all RGB benchmarks.

2   Related Work

RGB Visual Object Tracking: RGB Visual Object Tracking (VOT) has been a vibrant area of research in computer vision, with a rich history of developments and a diverse array of approaches. The goal in RGB VOT is to accurately recover the trajectory of a target object within a sequence of frames, often under various challenges such as occlusions, illumination changes, and scale variations.

The evolution of RGB VOT algorithms has seen a transition from traditional methods like Mean Shift [11] and CamShift [6] to more sophisticated approaches leveraging machine learning and deep learning. Correlation filters, such as the Discriminative Correlation Filter (DCF) [26] and its Kernelized version (KCF) [28], have significantly improved tracking performance by learning robust representations of the target object. Siamese networks, exemplified by SiamFC [3] and SiamRPN [37], have introduced a novel way to track objects by learning to match target appearances across frames. Discriminative models, including CSRDCF [17] and DiMP [4], have further enhanced tracking by distinguishing the target from the background using convolutional neural networks (CNNs). Lately, the introduction of Vision Transformers (ViTs) [20] has revolutionized the field, with models like Mixformer [12] and SeqTrack [7] demonstrating the ability to handle long-range dependencies and complex patterns, leading to state-of-the-art tracking performance on several benchmarks.

RGB-D Tracking:. RGB-D tracking represents a significant advancement in the realm of visual object tracking (VOT), capitalizing on the depth data provided by RGB-D sensors to bolster tracking accuracy. The Princeton Tracking Benchmark[50] has played a pivotal role in the evolution of this domain, establishing a robust platform for assessing both RGB and RGB-D tracking algorithms. This benchmark comprises 100 diverse RGB-D video datasets, encompassing a broad spectrum of tracking challenges, and has enabled a systematic evaluation of various tracking methodologies[51].

The integration of depth information has been proven to markedly enhance tracking capabilities, especially in situations of occlusions and model drift. The work of Ding et al.[18], have highlighted the benefits of incorporating depth images captured by devices like the Microsoft Kinect to refine conventional tracking techniques. The Depth Masked Discriminative Correlation Filter (DM-DCF)[33], introduced by Kart et al., exemplifies an RGB-D tracking algorithm that leverages depth segmentation for occlusion detection and dynamically adjusts the spatial support for correlation filters, achieving high performance in the Princeton RGBD Tracking Benchmark.

A recent trend is to fuse depth and infra red (IR) imaging information to build a multimodal tracker[67]. However, none of these methods can be directly applied to the RGB datasets.

Attention mechanisms:. The introduction of attention mechanism in computer vision aims to mimic the capability of allowing models to dynamically weight features based on their importance, thereby enhancing performance across a variety of visual tasks. Over time, attention mechanisms, categorized into channel, spatial, temporal, and branch attention [24], operate in various domains:

channel for feature selection, spatial for region focus, temporal for time-series analysis, and branch for selecting among model components.

Channel attention, like in Squeeze-and-Excitation (SE) Networks [29], boosts feature importance through channel weight recalibration. Spatial attention, exemplified by Deformable Convolutional Networks (Deformable ConvNets) [14], selectively focuses on regions of interest for enhanced feature representation. Temporal attention, as seen in the Temporal Adaptive Module (TAM) [41], captures intricate temporal relationships for improved video recognition. Branch attention, demonstrated in Highway Networks [52], enables dynamic information flow across layers, addressing the vanishing gradient problem in deep networks. In conclusion, attention mechanisms have become an indispensable tool in computer vision, offering a promising avenue for improving model performance and interpretability. As research continues, the integration of attention with other deep learning techniques and the development of more efficient and versatile attention models are expected to further advance the field.

In this work, we introduce a novel approach to RGB Tracking by integrating depth information estimated from the RGB images. Our approach addresses some limitations of the traditional RGB-only tracking. We achieve this integration through a novel depth attention mechanism, offering a depth prior for the tracking task without requiring expensive RGB-D cameras or retraining the model.
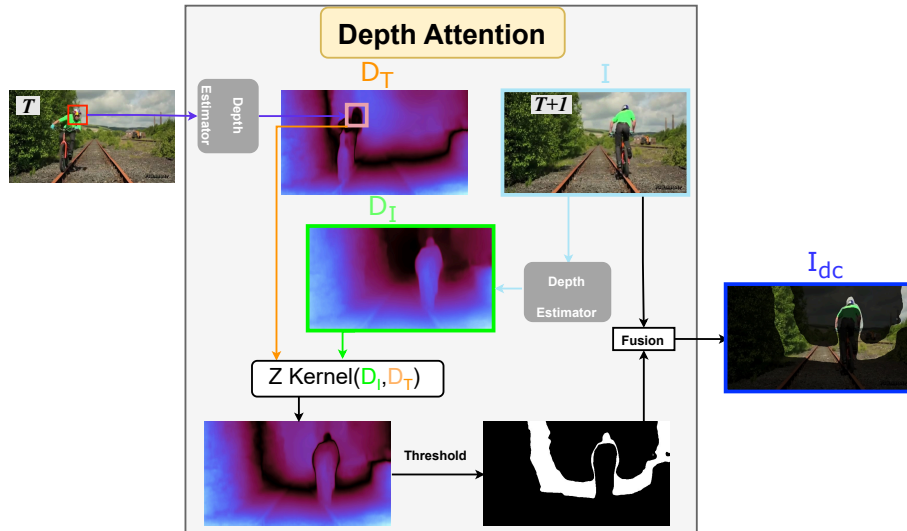

Fig. 2. Proposed depth attention to improve RGB tracking.

3 Depth Attention Based Tracking Framework

We present a simple framework at the core of which is a new module that utilizes a rapid monocular depth estimation algorithm to create an initial depth map. Towards addressing potential errors stemming from the estimation process, we further refine this depth information using a novel Z kernel, distinguishing our proposed method from the current RGB-D Tracking pipeline[62]. The outcome is a probability map serving as a valuable prior for the tracking algorithms. Notably, this seamless integration into the existing RGB tracking algorithms demonstrates its adaptability and wide applicability. Figure 2 sketches a schematic diagram of our proposed method.

3.1 Monocular Depth Estimation

Our monocular depth estimation approach leverages the Lite-Mono architecture [65], a lightweight hybrid of CNNs and Transformers for self-supervised learning. This method is tailored for applications where stereo or LiDAR data is not available, enabling depth estimation from single images. The Lite-Mono model employs Consecutive Dilated Convolutions (CDC) to capture multi-scale local features and Local-Global Features Interaction (LGFI) to encode global context, which is crucial for accurate depth inference. The training objective combines photometric reprojection loss and edge-aware smoothness loss to ensure visually coherent and spatially consistent depth maps. The photometric reprojection loss is defined as [65]:

$$L_r(\hat{I}_t, I_t) = \alpha/2 + (1 - \alpha) \cdot \|\hat{I}_t - I_t\| - \alpha SSIM(\hat{I}_t, I_t)/2 \qquad (1)$$

where α is typically set to 0.85, and $SSIM$ denotes the Structural Similarity Index. The edge-aware smoothness loss encourages spatial consistency in the predicted depth map d_t as follows [65]:

$$L_{smooth} = \left| \frac{\partial d_t^*}{\partial x} \right| e^{-|\frac{\partial I_t}{\partial x}|} + \left| \frac{\partial d_t^*}{\partial y} \right| e^{-|\frac{\partial I_t}{\partial y}|}, \qquad (2)$$

where $d_t^* = \frac{d_t}{\bar{d}_t}$ is the mean-normalized inverse depth, and $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ represent the spatial gradients. The total loss function is a weighted sum of these two losses:

$$L = \frac{1}{3} \sum_{s \in \{1, 1/2, 1/4\}} (L_r + \lambda L_{smooth}), \qquad (3)$$

where s is the output of depth decoder at different scales, and λ is a hyperparameter. This approach allows for efficient and accurate depth estimation, making it suitable for real-time applications.

3.2 Z Kernel and Signal Modulation

Given a single RGB image, monocular depth estimation provides a single-channel score map, where each pixel value represents an estimated depth. However, this

raw depth map is not directly suitable for existing RGB tracking algorithms, as these algorithms typically operate with input channels limited to 3. To address this, we propose our custom Z_K kernel to create a probability map that highlights the region of interest within the bounding box:

$$Z_K(D_I, D_T) = \left| \frac{(D_I - \mathrm{ME}(D_T))}{\mathrm{ME}(|D_T - \mathrm{ME}(D_T)|)} \right|, \tag{4}$$

where D_I is the depth matrix of the entire current frame, while D_T denotes the depth matrix confined to the bounding box of the previous frame. The function $\mathrm{ME}(.)$ computes the median of the depth matrix. The Z_K kernel measures the distance of D_I and D_T. By setting a threshold Th on Z_K, a mask is obtained, focusing on the region of interest. Then we modulate the image with a linear formula:

$$I_{dc} = (1 - k_1)I + k_1 I \odot \mathrm{Mask}(Z_K, Th), \tag{5}$$

where $\odot$ is element wise multiplication operator, and k_1 is the depth attention confidence. I_{dc} represents the modulated image, the function $\mathrm{Mask}(Z_K, th)$ transforms the depth map Z_K into a binary mask based on the threshold th. Balancing computational efficiency and precision, the D_T undergoes periodic updates, specifically every 60 frames. Our method's periodic updates are controlled by statistical analysis, which indicates that across six key benchmarks, the target's movement is relatively small within certain frames, as further discussed in Section 4. This insight anchors our strategy, ensuring that our approach maintains both adaptability and precision in the ever-changing landscape of visual tracking. The calculation for determining k_1 follows the formula:

$$k_1 = \frac{p(v_n; \mu, \sigma)}{\sum_{i=1}^n p(v_i; \mu, \sigma)}, \tag{6}$$

where v_i denotes the peak-to-sidelobe ratio (PSR) of the confidence map, $p(\cdot)$ represents the Gaussian probability density function with parameters μ and σ. This formula is designed to calculate the proportion of the masked image. Instead of simply using the masked image on its own, we discovered that blending it with the original image significantly enhances the overall performance of our system. We would further verify this in Section 4. μ and σ are estimated from the collected PSR values $v_i, i = 1, 2, 3, ..., n$. In the code, we specifically set $n = 5$. For better understanding, we draw the results in Figure 3. This entire procedure can be conceptualized as a disentangling process for the target object from the background. Moreover, note that this method is compatible with all existing tracking methods.

4   Experiments

Datasets: Our approach has been rigorously validated on several benchmark datasets, which are widely recognized for their role in assessing the performance of visual tracking algorithms.
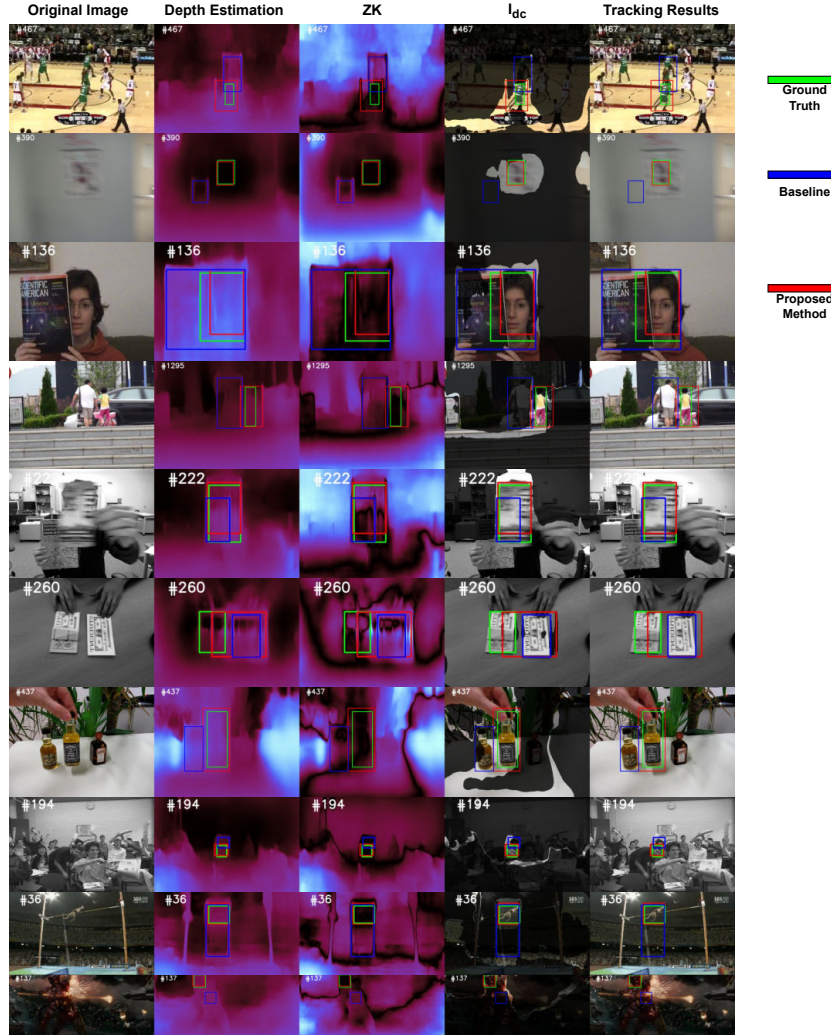
Fig. 3. This figure shows the original images, their depth estimations, Z_K values, I_{dc}, and tracking outcomes across four different sequences. Noticeably, when the baseline tracker starts to drift, our proposed method effectively prevents this by generating a mask with zero values in the background area. The proposed depth attention based masking enhances the overall tracking performance.

– **OTB100** [58]: This dataset consists of 100 video sequences with a diverse range of tracking challenges, such as illumination changes, occlusions, and

camera motion. It is one of the most popular benchmarks for evaluating tracking algorithms due to its comprehensiveness and difficulty.

- **NfS** [34]: The Need for Speed (NfS) dataset is designed to test tracking algorithms under high frame rates, capturing videos with high-speed motion and rapid scene changes. It pushes the boundaries of tracking algorithms by introducing extreme tracking scenarios.
- **AVisT** [47]: The AVisT dataset focuses on tracking under adverse visibility conditions, such as low-light and high clutter. It presents a significant challenge for trackers, as it requires robust tracking in scenarios where visual information is limited.
- **UAV123** [45]: This dataset is specifically tailored for tracking small objects in aerial videos, often at high altitudes and with significant perspective changes. It introduces unique challenges due to the scale and motion of the objects in the context of aerial imagery.
- **LaSOT** [21]: The LaSOT dataset is a large-scale single object tracking benchmark that includes a diverse set of 1,550 videos. It is designed to provide a comprehensive evaluation of tracking algorithms across various challenging conditions, such as occlusions, scale variations, and fast motion.
- **GOT-10k** [30]: The GOT-10k dataset is a large-scale benchmark for generic object tracking in the wild, containing over 10,000 video clips with a wide variety of object classes and tracking scenarios. It is known for its high diversity and the complexity of the tracking tasks it presents.
- **NT-VOT211** [43]: This dataset represents a novel benchmark, specifically designed for evaluating visual object tracking algorithms under challenging low-light conditions. It comprises 211 videos and 211k annotated frames. It poses the unique challenges inherent to the nighttime tracking scenarios.

Implementation Details: All results in this paper were obtained within a consistent computational environment, utilizing Python 3.8.10, PyTorch 1.11.0 with CUDA 11.3, NumPy 1.22.3, and OpenCV 4.8.0. The Depth Estimation model was initially trained following the procedures outlined in [65]. The training process was conducted on a single Tesla V100 GPU, while subsequent evaluations were performed using a Nvidia A5000 GPU. Here, the sole hyperparameter Th in equation 5 was maintained at a constant value of $Th = 1.5$.

Evaluation metrics: We use a range of metrics, including AUC, success scores, and AO, tailored to each benchmark as specified in their respective papers[58,21,30,34]. NT-VOT211 dataset[43] adheres to the evaluation protocol of LaSOT[21], ensuring a consistent and fair assessment of tracking performance.

Results: We integrate the proposed method into ODTrack [66], the SOTA on LaSOT [21], TrackingNet [46], TNL2k [54], and ITB [38]. We have listed all the results in Table 1. As shown, we have achieved new SOTA on all six benchmarks. The better performance highlighted in **red bold font**, and '+DA' indicates the integration of proposed Depth Attention (DA) module.

Next, we integrate the proposed Depth Attention (DA) module into 2023 SOTA and evaluate on different benchmarks. We have listed all the results in Table 2. We observe a consistent performance improvement highlighted in **red bold font**.

Table 1. Comparison of state-of-the-art methods with and without the proposed Depth Attention (DA) module on six benchmarks: GOT10K, LaSOT, TrackingNet, TNL2k, UAV123, and NfS. The best results are highlighted in **red** and the second best in **blue** colors.

Method	GOT10k[30] (AO)	LaSOT[21] (AUC)	TrackingNet[46] (Success)	TNL2k[35] (AUC)	UAV123[45] (AUC)	Nfs[34] (AUC)
TransT[9]	64.9	67.1	81.4	50.7	-	-
OSTrack[64]	73.7	71.1	83.9	55.9	-	-
SwinTrack[39]	72.4	71.3	84.0	-	-	-
Mixformer-22k[13]	70.7	69.2	83.1	-	69.5	-
SeqTrack-B_{384}[8]	74.5	71.5	83.9	-	68.6	66.7
VideoTrack[59]	72.9	70.2	83.8	-	-	-
RTS[49]	73.5	69.7	81.6	54.7	67.6	65.4
RTS+DA (31 fps)	74.2	70.3	82.1	55.6	68.1	66.0
DropTrack[57]	75.9	71.8	84.1	57.7	69.5	66.7
DropTrack+DA (22 fps)	77.0	72.4	84.6	58.2	70.4	67.2
ARTrack[56]	76.6	70.4	84.0	57.5	67.7	66.8
ARTrack+DA (21 fps)	77.5	70.7	84.2	57.7	68.1	67.4
ODTrack[66]	77.0	73.4	85.7	61.7	69.5	67.6
ODTrack+DA (8fps)	78.8	74.3	86.1	61.9	70.3	68.5

Comparison with SOTA RGBD tracking methods: First, we evaluate VIPT[67] (RGBD tracker) and ODTrack + DA (ours) on GOT10k. ODTrack[66] is a recent SOTA on many datasets. To incorporate depth information into VIPT, a necessity for RGBD tracking algorithms, we utilize depth estimates from a monocular architecture to estimate the depth information of GOT-10k video. Second, we compare VIPT and VIPT + DA (ours) on DepthTrack dataset[61], a RGBD dataset.

Ablation study on k_1: As stated in Equation 5, the confidence of the depth attention mechanism significantly influences the extent to which the original image is modulated. We conducted experiments by varying the parameter k_1 across different constants in the RTS method [49]. See Tables 4 and Table 5.

Subsequently, we applied the optimal constant $k_1 = 0.02$ to the module and compared it with the proposed module, where k_1 is estimated using Equation 6. The results are illustrated in Figure 4. It is evident that the adaptive estimation process for k_1 plays a vital role in enhancing performance.

Qualitative Results: Figure 5, illustrates the tracking process of KeepTrack[44] and KeepTrack with our depth attention. Our proposed method is adept at handling challenging scenarios e.g., occlusion, motion blur, and camera motion.

Attributes-wise Evaluation: In addition, we assessed KeepTrack+DA and other top trackers in NT-VOT211[43], with the results illustrated in Fig 6. This is attributes-wise evaluation. The results demonstrate that the proposed method excels in challenging scenes, such as out-of-view, motion blur, and fast motion. It tends to maintain original performance on camera motion, deformation, occlusion and distractor scenarios, while potentially exhibiting reduced performance in tiny target scenarios.

Frequency Analysis: Inspired by the work of park et al. [48], we selected a Resnet-50[25] and a ViT Tiny Patch 16/224[19] and applied the proposed method to both. Subsequently, we conducted Fourier analysis on these models, with the results depicted in Figure 7. It is evident from the analysis that the proposed
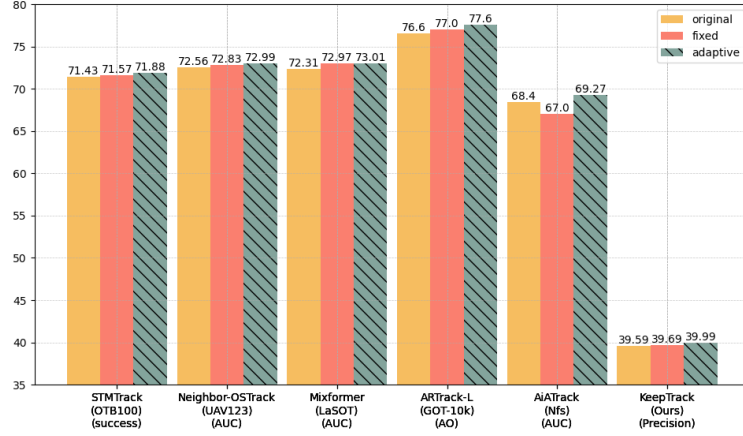
Fig. 4. We compare the original algorithm, the depth attention module with a fixed hyperparameter (Th=1.5, $k_1 = 0.02$), and the depth attention module with an adaptive hyperparameter on k_1. We elaborate on the base algorithm, dataset, and the metrics used along the axes.
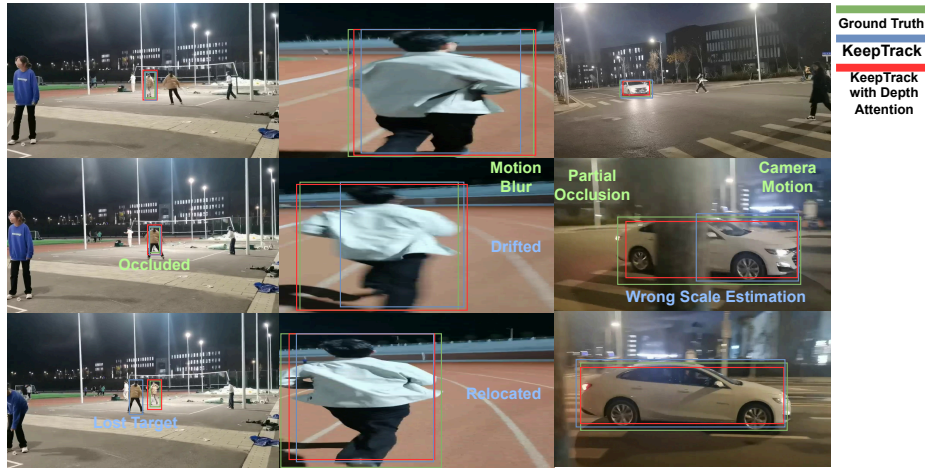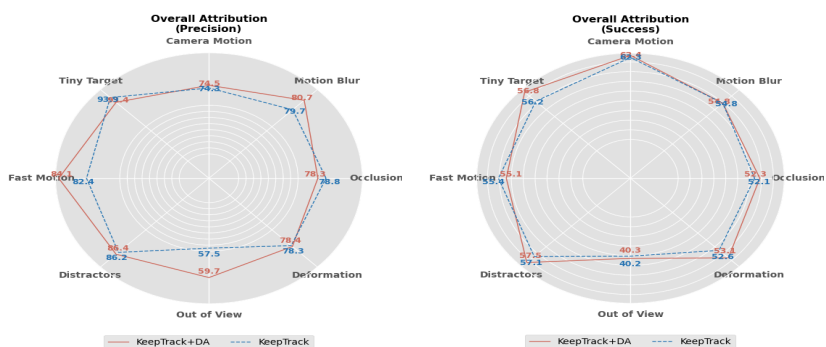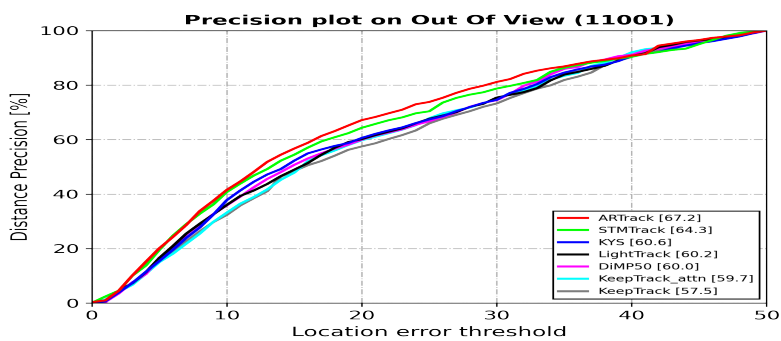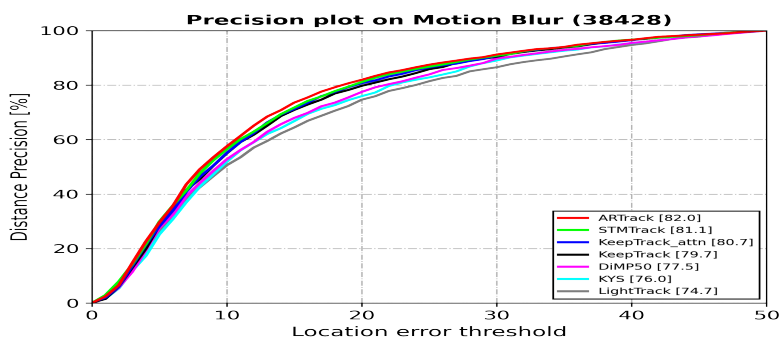

Fig. 5. Visual comparison between KeepTrack and KeepTrack+DA (ours). It can be seen that, our method is robust under the challenging scenarios of occlusion and motion blur.

(a)

(b)

Fig. 6. Evaluation of attributes on NT-VOT211. In subfigure (a), we display the attribution radar graph, comparing the original model with the model incorporating depth attention. Subfigure (b) features a detailed plot, highlighting the attributes where our module showed the most improvement, specifically Motion Blur and Out-of-View.

Table 2. More comparisons with SOTA methods. Addition of Depth Attention (DA) module has consistently improved performance. The evaluation measures Average Overlap (AO), Success (S), Precision (P), Normalized Precision (NP) are used; OP50 refers to the Overlap Precision at 50% threshold and OP75 refers to the Overlap Precision at 75% threshold, SR50 refers to the Success Rate at a 50% IoU threshold, SR75 refers to the Success Rate at a 75% IoU threshold. The 'rank' column denotes the ranking of each method on their corresponding datasets, with the specific metric used for ranking provided in parentheses.

Method	Dataset	rank	Metric				
ARTrack-L [56](24 fps)	GOT-10k [30]	3(AO)	AO 76.6	SR50 85.6	SR75 75.4		
ARTrack-L+**DA**(21 fps)		2(AO)	**77.6**	**87.2**	**76.6**		
STMTrack [22](43 fps)	OTB100[58]	2(S)	S 71.43			P 92.64	
STMTrack+**DA**(39 fps)		1(S)	**71.88**			**93.01**	
Neighbor-OSTrack [10](53 fps)	UAV123 [45]	2(AUC)	AUC 72.56	P 93.37	NP 88.51	OP50 87.75	OP75 68.15
Neighbor-OSTrack+**DA**(49 fps)		1(AUC)	**72.99**	**94.12**	**89.00**	**89.01**	**68.78**
Mixformer(ConvMAE)-L [13](27 fps)	LaSOT [21]	5(AUC)	AUC 72.31	P 78.93	NP 81.66	OP50 83.63	OP75 72.17
Mixformer(ConvMAE)-L+**DA**(25 fps)		3(AUC)	**73.01**	**80.83**	**83.57**	**84.39**	**73.87**
AiATrack [23](35 fps)	NfS [34]	2(AUC)	AUC 68.40	P 84.70	NP 87.29	OP50 87.37	OP75 **55.08**
AiATrack+**DA**(33 fps)		1(AUC)	**69.27**	**85.38**	**88.98**	**88.84**	55.05
KeepTrack [44](19 fps)	NT-VOT211 [43]	2(P)	AUC 39.59	P 55.50	NP 85.06	OP50 50.52	OP75 **12.83**
KeepTrack+**DA**(15 fps)		1(P)	**39.99**	**55.72**	**85.88**	**50.90**	12.77

Table 3. Comparison with RGBD trackers.

Strategy 1	Method		Strategy 2	Method	
GOT10k	VIPT	72.1	DepthTrack	VIPT	57.8
(AO)	ODTrack+DA	**78.8**	(F scores)	VIPT+DA	**58.6**

method exhibits greater similarity to the Conv-based approach in terms of frequency analysis. In the initial segments of the network, the proposed method demonstrates minimal differences, except in the vicinity around 0.2π. However, as the network progresses deeper, these distinctions become more pronounced, contributing to enhanced performance. Notably, for multi-head self-attentions (MSAs), the impact of incorporating the proposed method is particularly prominent within the interval $[0.1\pi, 0.8\pi]$, and this effect intensifies with increasing network depth.

The observed trend in deeper networks indicates that the proposed method tends to rectify high-frequency components with larger amplitudes. In essence, the proposed method aims to attenuate the processing of the ViT-based method, preventing it from acting as a low-pass filter in the context of frequency analysis, as discussed in [48]. This observation is further supported by the fact that Conv-based methods, specifically STMTrack and KeepTrack, exhibit the least improvement compared to ViT-based trackers.

Table 4. Ablation study on k_1, We evaluated the RTS under different parameters on OTB100

Hyper-parameter k_1 (Th=1.5)	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.02
AUC	34.60	56.47	59.13	61.55	63.14	64.87	64.40	64.60	65.45	66.09	66.17	**66.35**
Precision	43.63	75.15	79.14	82.85	84.66	87.02	85.92	86.04	87.10	88.14	87.94	**88.45**
Norm Precision	38.38	65.41	69.11	72.82	74.29	76.52	75.93	75.77	76.75	77.74	77.66	**78.36**

Table 5. Optimal fixed k_1 and Th

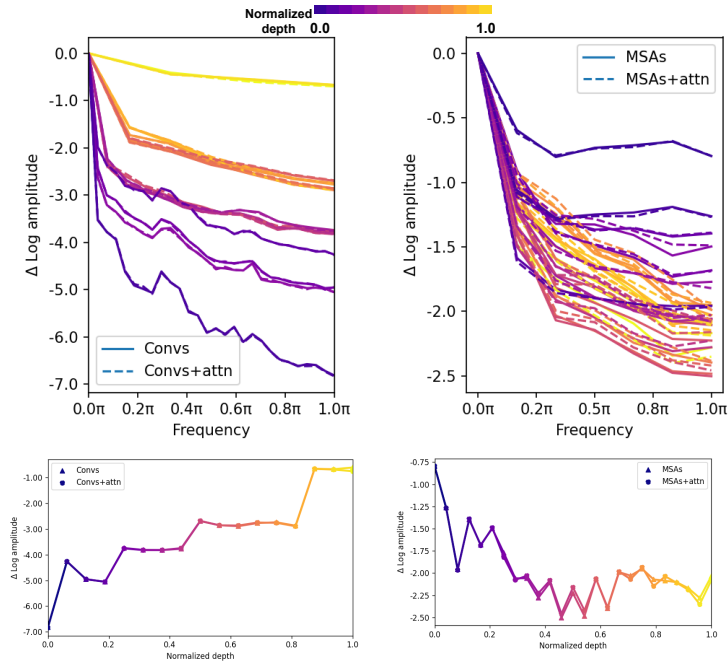| Hyper-parameter (Th=1.5) (k_1=0.02)	OTB		LaSOT		AVisT			GOT-10k	
	RTS	RTS+DA	RTS	RTS+DA	RTS	RTS+DA		RTS	RTS+DA
AUC	66.17	**66.35**	69.46	**70.04**	49.98	**50.81**	AO	73.5	74.1
Precision	87.94	**88.45**	73.21	**73.89**	47.78	**48.74**	$SR_{0.5}$	83.1	**83.7**
Norm Precision	77.6	**78.36**	76.05	**76.83**	65.46	**66.12**	$SR_{0.75}$	68.3	**68.7**

Fig. 7. First row: frequency component analysis, where different depths are represented by distinct colors. A depth-indicating bar is provided at the top for reference. Second row: we present statistics on the overall amplitude corresponding to different depths.
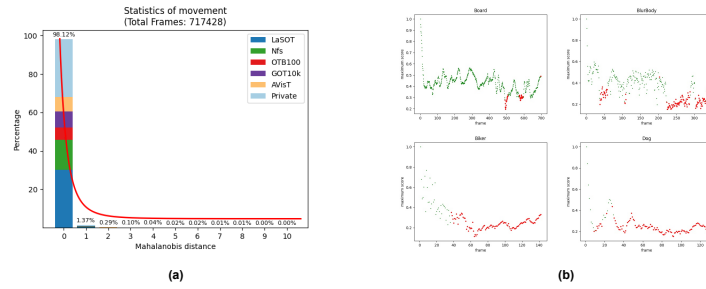
Fig. 8. In subfigure a, we analyzed Manhattan Distance displacement across six benchmarks, totaling 717,428 frames. For each set of five consecutive frames, we measured the target's movement relative to its own dimensions (one unit representing length or width). Findings indicate that, in most cases, the target remained stationary, with movements not exceeding its dimensions. In subfigure b, we computed the Peak-to-Sidelobe Ratio (PSR) of the score map generated by Tracker STMTrack, showing exponential decay within a time interval. Though the algorithm may correct this in subsequent intervals, the deteriorating PSR value ultimately leads to tracker failure, marked by the red point. This highlights the need for a periodically updated prior.

Why every five frames? Through statistical analysis of datasets including Nfs[34], AVisT[47], LaSOT[21], OTB100[58], GOT-10k[30], and NT-VOT211, we identified a distinct motion pattern for the target, as illustrated in part a of Figure 8. This pattern exhibits a long-tail distribution. This can be mathematically expressed as $\lim_{x \to 0} P(|m - x| < \epsilon) = 1$. Where 'm' represents the movement, and it suggests that as the degree of movement within the 5 frames decreases, the likelihood of the statement being true increases. This ensures that the subsequent 4 frames can be roughly regarded as the Value Component associated with the initial frame.

5   Conclusion and Limitations

Our work presents a novel approach to Visual Object Tracking (VOT) by integrating monocular depth estimation in a principled way, resulting in a robust tracking system capable of handling occlusions and motion blur. We note state-of-the-art performance of our method on multiple benchmarks, including OTB100, UAV123, LaSOT, and GOT-10k. The ablation study and Fourier analysis confirm the effectiveness of our approach in overcoming occlusions and maintaining tracking continuity. However, the effectiveness of depth attention could be further enhanced through an end-to-end training approach, which is an interesting avenue for future work.

References

1. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.: Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1401–1409 (2016)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: 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. pp. 850–865. Springer (2016)
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II pp. 850–865 (2016)
4. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. Proceedings of the IEEE/CVF international conference on computer vision pp. 6182–6191 (2019)
5. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 2544–2550. IEEE (2010)
6. Bradski, G.R., Kaehler, A., Willis, C., Huang, J., Pereira, B., Nister, D., Koller, D., Chellappa, R.: Computer vision face tracking for use as a human computer interface. Computer **33**(2), 49–58 (2000)
7. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: Sequence to sequence learning for visual object tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 14572–14581 (2023)
8. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: Sequence to sequence learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14572–14581 (2023)
9. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8126–8135 (2021)
10. Chen, Y.H., Wang, C.Y., Yang, C.Y., Chang, H.S., Lin, Y.L., Chuang, Y.Y., Liao, H.Y.M.: Neighbortrack: Single object tracking by bipartite matching with neighbor tracklets and its applications to sports. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5138–5147 (June 2023)
11. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5), 603–619 (2003)
12. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 13608–13618 (2022)
13. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13608–13618 (2022)
14. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H.: Deformable convolutional networks. Proceedings of the IEEE International Conference on Computer Vision pp. 764–773 (2017)
15. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6638–6646 (2017)

16. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British machine vision conference, Nottingham, September 1-5, 2014. Bmva Press (2014)

17. Danelljan, M., Shahbaz Khan, F., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 1090–1097 (2014)

18. Ding, P., Song, Y.: Robust object tracking using color and depth images with a depth based occlusion handling and recovery. In: 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). pp. 1–6. IEEE (2015)

19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. 9th International Conference on Learning Representations, ICLR 2021 (2021), `https://openreview.net/forum?id=YicbFdNTTy`

21. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)

22. Fu, Z., Liu, Q., Fu, Z., Wang, Y.: Stmtrack: Template-free visual tracking with space-time memory networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13774–13783 (2021)

23. Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J.: Aiatrack: Attention in attention for transformer visual tracking. In: European Conference on Computer Vision. pp. 146–164. Springer (2022)

24. Guo, M.H., Xu, T.X., Liu, J.J., et al.: Attention mechanisms in computer vision: A survey. Computational Visual Media **8**(3), 331–368 (2022)

25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

26. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV pp. 702–715 (2012)

27. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence **37**(3), 583–596 (2014)

28. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2014)

29. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 2107–2121 (2018)

30. Huang, L., Zhao, X., Huang, K.: 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 (2019)

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

32. Kart, U., Lukezic, A., Kristan, M., Kamarainen, J.K., Matas, J.: Object tracking by reconstruction with view-specific discriminative correlation filters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1339–1348 (2019)

33. Kart, U., Kämäräinen, J.K., Matas, J., Cricri, F.: Depth masked discriminative correlation filter. In: arXiv preprint arXiv:1804.05646 (2018)

34. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1125–1134 (2017)

35. Kim, M., Lee, S., Ok, J., Han, B., Cho, M.: Towards sequence-level training for visual tracking. In: European Conference on Computer Vision. pp. 534–551. Springer (2022)

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

37. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 8971–8980 (2018)

38. Li, X., Liu, Q., Pei, W., Shen, Q., Wang, Y., Lu, H., Yang, M.H.: An informative tracking benchmark. arXiv preprint arXiv:2112.06467 (2021)

39. Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022),

40. Liu, Y., Jing, X.Y., Nie, J., Gao, H., Liu, J., Jiang, G.P.: Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos. IEEE Transactions on Multimedia **21**(3), 664–677 (2018)

41. Liu, Z., Wang, L., Wu, W., Qian, C.: Tam: Temporal adaptive module for video recognition. arXiv preprint arXiv:2005.06803 (2020)

42. Lukezic, A., Matas, J., Kristan, M.: D3s-a discriminative single shot segmentation tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7133–7142 (2020)

43. Yu, Liu., Mehmood, A., Khan, M.H.: NT-VOT211: A Large-Scale Benchmark for Night-time Visual Object Tracking. In: Proceedings of Asian Conference on Computer Vision (ACCV). (2024)

44. Mayer, C., Danelljan, M., Paudel, D.P., Van Gool, L.: Learning target candidate association to keep track of what not to track. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13444–13454 (2021)

45. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 445–461. Springer (2016)

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

47. Noman, M., Ghallabi, W.A., Kareem, D., Mayer, C., Dudhane, A., Danelljan, M., Cholakkal, H., Khan, S., Gool, L.V., Khan, F.S.: Avist: A benchmark for visual object tracking in adverse visibility. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. p. 817. BMVA Press (2022), `https://bmvc2022.mpi-inf.mpg.de/817/`

48. Park, N., Kim, S.: How do vision transformers work? arXiv preprint arXiv:2202.06709 (2022)

49. Paul, M., Danelljan, M., Mayer, C., Van Gool, L.: Robust visual tracking by segmentation. In: European Conference on Computer Vision. pp. 571–588. Springer (2022)

50. Song, S., Xiao, J.: Tracking revisited using rgbd camera: Baseline and benchmark. In: arXiv preprint arXiv:1204.06302 (2012)

51. Song, S., Xiao, J.: Tracking revisited using rgbd camera: Unified benchmark and baselines. In: 2013 IEEE International Conference on Computer Vision (ICCV). pp. 3546–3553. IEEE (2013)

52. Srivastava, R., Greff, K., Schmidhuber, J.: Highway networks pp. 2377–2385 (2015)

53. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 1328–1338 (2019)

54. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 13763–13773. Computer Vision Foundation / IEEE (2021).

55. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9697–9706 (June 2023)

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

57. Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B.: Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14561–14571 (2023)

58. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015). `https://doi.org/10.1109/TPAMI.2014.2388226`

59. Xie, F., Chu, L., Li, J., Lu, Y., Ma, C.: Videotrack: Learning to track objects via video transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 22826–22835. IEEE (2023). `https://doi.org/10.1109/CVPR52729.2023.02186`, `https://doi.org/10.1109/CVPR52729.2023.02186`

60. Yan, S., Yang, J., Kapyla, J., Zheng, F., Leonardis, A., Kamarainen, J.K.: Depthtrack: Unveiling the power of rgbd tracking. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

61. Yan, S., Yang, J., Käpylä, J., Zheng, F., Leonardis, A., Kämäräinen, J.: Depthtrack: Unveiling the power of RGBD tracking. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 10705–10713. IEEE (2021). `https://doi.org/10.1109/ICCV48922.2021.01055`, `https://doi.org/10.1109/ICCV48922.2021.01055`

62. Yang, J., Gao, S., Li, Z., Zheng, F., Leonardis, A.: Resource-efficient rgbd aerial tracking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
63. Yang, J., Li, Z., Yan, S., Zheng, F., Leonardis, A., Kamarainen, J.K., Shao, L.: Rgbd object tracking: An in-depth review. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
64. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European Conference on Computer Vision. pp. 341–357. Springer (2022)
65. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18537–18546 (June 2023)
66. Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X.: Odtrack: Online dense temporal token learning for visual tracking. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 7588–7596. AAAI Press (2024). https://doi.org/10.1609/AAAI.V38I7.28591, https://doi.org/10.1609/aaai.v38i7.28591
67. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 9516–9526. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00918, https://doi.org/10.1109/CVPR52729.2023.00918
68. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 101–117 (2018)