

# Supplementary Materials: Ego3DT: Tracking All 3D Objects in Ego-Centric Video of Daily Activities

Anonymous Authors

The supplementary material is structured as follows:

- We begin with the “Ego3DT Pseudo Code” section, detailing the **Ego3DT** method, encompassing input requirements, variables, functions, and the procedure of the **Ego3DT** Method algorithm in Section 1.
- We provide the “Detailed Performance Comparison” section, including details of our results with other comparison methods.
- We present the “Temporal Windows Impact” section to illustrate the impact of the time windows on tracking performance.
- We show the “Demo Visualization” section to visualize the tracking and the 3D matching.

## 1 EGO3DT PSEUDO CODE

The **Ego3DT** framework is predicated on the concept of dynamic 3D scene construction and object trajectory estimation from ego-centric video data. Using advanced models for object detection, segmentation, and 3D position estimation, Ego3DT processes temporal frames from a single viewpoint, assigning global IDs to objects and outputting their trajectories based on precise 3D positioning. Here, we present the pseudo-code that encapsulates this process.

As shown in Algorithm A1 the method starts by initializing the required models for 2D detection, semantic segmentation, and 3D estimation. These models work together for each frame in the video sequence to detect and segment objects in 2D before estimating their 3D coordinates. During the initial phase within the first-time window, a combination of KDTree distance computation and the Hungarian algorithm is employed to assign global IDs to objects, ensuring the correct tracking of each object as they move through space and time.

Subsequent frames are processed using a sliding window approach. The framework updates the 3D scene for each new frame by aligning it with the previous frame’s data using a 3D scene alignment method. The aligned data is then used to perform point matching, crucial for maintaining object continuity across different temporal windows. This iterative process results in a buffer that contains the tracking information for all processed frames.

## 2 DETAILED PERFORMANCE COMPARISON

We show the details of each scene in Table B1. We have collected six scenes for experiments containing indoor (Market,

### Algorithm A1 Ego3DT Method

```
1: Input: Ego-centric video frames  $X = \{I_i\}_{i=1}^N$ 
2: Output: Tracked objects  $Y$  with 3D coordinates and IDs
3: Initialize models: 2D Object Detector Det, 2D Segmenter Seg, 3D Estimator  $\mathcal{G}$ 
4: for each frame  $I_t$  in video frames  $X$  do
5:    $O_{2D}^t \leftarrow \text{Det}(I_t)$  {Detect objects in 2D}
6:    $O_{Seg}^t \leftarrow \text{Seg}(O_{2D}^t, I_t)$  {Segment objects}
7:    $O_{3D}^t \leftarrow \mathcal{G}(O_{Seg}^t)$  {Estimate 3D coordinates}
8: end for
9: Initialize: Step size  $S = W - T$ , Buffer  $\mathcal{B} \leftarrow \emptyset$ 
10:  $Y_0 \leftarrow \text{Hungarian}(\text{PointMatch}(O_{3D}^1))$ 
11: Add  $Y_0$  to  $\mathcal{B}$ 
12: for  $t = W + 1$  to  $N$  step  $S$  do
13:    $O_{3D}^t \leftarrow \mathcal{G}(X, \text{Seg}(\text{Det}(I_t)))$ 
14:   Align 3D scenes:  $O_{3D}^t \leftarrow \mathcal{A}(O_{3D}^{t-S}, O_{3D}^t)$ 
15:    $Y_t \leftarrow \text{PointMatch}(O_{3D}^{t-S}, O_{3D}^t)$ 
16:   Add  $Y_t$  with IDs to  $\mathcal{B}$ 
17: end for
18: Convert buffer  $\mathcal{B}$  to the output space  $Y$ 
19: return  $Y$ 
```

### Algorithm A2 Objects Tracking in First Window

```
1: Input: Initial 3D coordinates  $O_{3D}^1$ , Window size  $\mathcal{W}$ 
2: Output: Tracked objects  $Y_0$ 
3: for each pair of frames  $(i, i + 1)$  where  $i \leq \mathcal{W} - 1$  do
4:   Compute KDTree distance between all points in  $O_{3D}^i$  and  $O_{3D}^{i+1}$ 
5:   Apply Hungarian algorithm using distances as costs
6:   Assign IDs based on matches to obtain  $Y_i$ 
7: end for
8: Aggregate all  $Y_i$  to form  $Y_0$ 
9: return  $Y_0$ 
```

Corridor, and Kitchen) and outdoor scenes (Garden1, Garden2, and Garage). From the results, our method outperforms other existing trackers.

## 3 TEMPORAL WINDOWS IMPACT

The temporal extent of the tracking window, or temporal window, plays a critical role in the efficacy of object tracking within the **Ego3DT** framework. The choice of window size directly influences the tracking accuracy, the consistency of object identification, and the computational efficiency of

Table B1: Comparison of Open Vocabulary MOT performance.

Scene	Market			Garden1		
Tracker	HOTA (↑)	IDF1 (↑)	DetA (↑)	HOTA (↑)	IDF1 (↑)	DetA (↑)
ByteTrack [6] + YOLO-World [1]	16.95	15.92	16.26	13.01	14.21	15.87
ByteTrack [6] + GLEE [5]	29.87	<b>30.97</b>	35.82	16.09	19.61	20.03
DeepSort [4] + YOLO-World [1]	9.84	8.86	10.49	7.94	7.37	9.38
DeepSort [4] + GLEE [5]	15.82	16.19	21.34	10.78	11.86	14.26
OVTrack [3]	15.00	14.35	15.14	10.54	10.32	8.93
TET [2]	13.65	12.84	13.90	12.31	11.31	9.00
<b>Ego3DT (Ours) + OVTrack [3]</b>	12.85	11.46	20.44	3.24	1.65	9.14
<b>Ego3DT (Ours) + TET [2]</b>	9.12	9.24	15.67	11.45	11.98	10.62
<b>Ego3DT (Ours) + YOLO-World [1]</b>	17.10	15.61	18.10	18.33	16.76	24.54
<b>Ego3DT (Ours) + GLEE [5]</b>	<b>30.31</b>	28.03	<b>54.52</b>	<b>27.47</b>	<b>26.86</b>	<b>47.15</b>
Scene	Corridor			Garden2		
Tracker	HOTA (↑)	IDF1 (↑)	DetA (↑)	HOTA (↑)	IDF1 (↑)	DetA (↑)
ByteTrack [6] + YOLO-World [1]	26.92	29.25	15.85	15.69	18.79	19.62
ByteTrack [6] + GLEE [5]	<b>41.03</b>	<b>52.34</b>	31.96	25.43	<b>30.16</b>	28.63
DeepSort [4] + YOLO-World [1]	11.78	10.92	8.49	13.53	18.31	14.94
DeepSort [4] + GLEE [5]	20.41	20.43	17.99	19.15	22.99	20.97
OVTrack [3]	18.66	18.83	11.47	10.96	12.28	13.84
TET [2]	20.74	22.07	10.75	6.91	8.60	8.63
<b>Ego3DT (Ours) + OVTrack [3]</b>	20.13	20.41	12.77	10.50	15.17	16.42
<b>Ego3DT (Ours) + TET [2]</b>	19.84	17.70	12.43	6.84	9.83	9.58
<b>Ego3DT (Ours) + YOLO-World [1]</b>	22.53	20.23	15.97	18.30	21.23	25.22
<b>Ego3DT (Ours) + GLEE [5]</b>	37.64	38.19	<b>33.49</b>	<b>27.67</b>	27.60	<b>50.42</b>
Scene	Garage			Kitchen		
Tracker	HOTA (↑)	IDF1 (↑)	DetA (↑)	HOTA (↑)	IDF1 (↑)	DetA (↑)
ByteTrack [6] + YOLO-World [1]	18.55	20.34	13.33	15.26	12.76	20.28
ByteTrack [6] + GLEE [5]	31.84	28.61	18.91	27.49	<b>25.67</b>	34.11
DeepSort [4] + YOLO-World [1]	10.22	9.62	12.26	9.96	7.28	14.50
DeepSort [4] + GLEE [5]	12.34	12.80	12.21	14.34	13.03	19.58
OVTrack [3]	11.53	8.48	22.75	19.88	22.37	20.27
TET [2]	12.93	9.80	6.44	11.52	10.48	15.64
<b>Ego3DT (Ours) + OVTrack [3]</b>	7.88	8.39	13.88	15.75	18.51	21.81
<b>Ego3DT (Ours) + TET [2]</b>	9.44	7.17	8.62	9.77	10.17	18.11
<b>Ego3DT (Ours) + YOLO-World [1]</b>	12.29	13.11	12.34	14.56	12.65	22.34
<b>Ego3DT (Ours) + GLEE [5]</b>	<b>33.70</b>	<b>40.28</b>	<b>34.60</b>	<b>28.19</b>	25.54	<b>59.34</b>

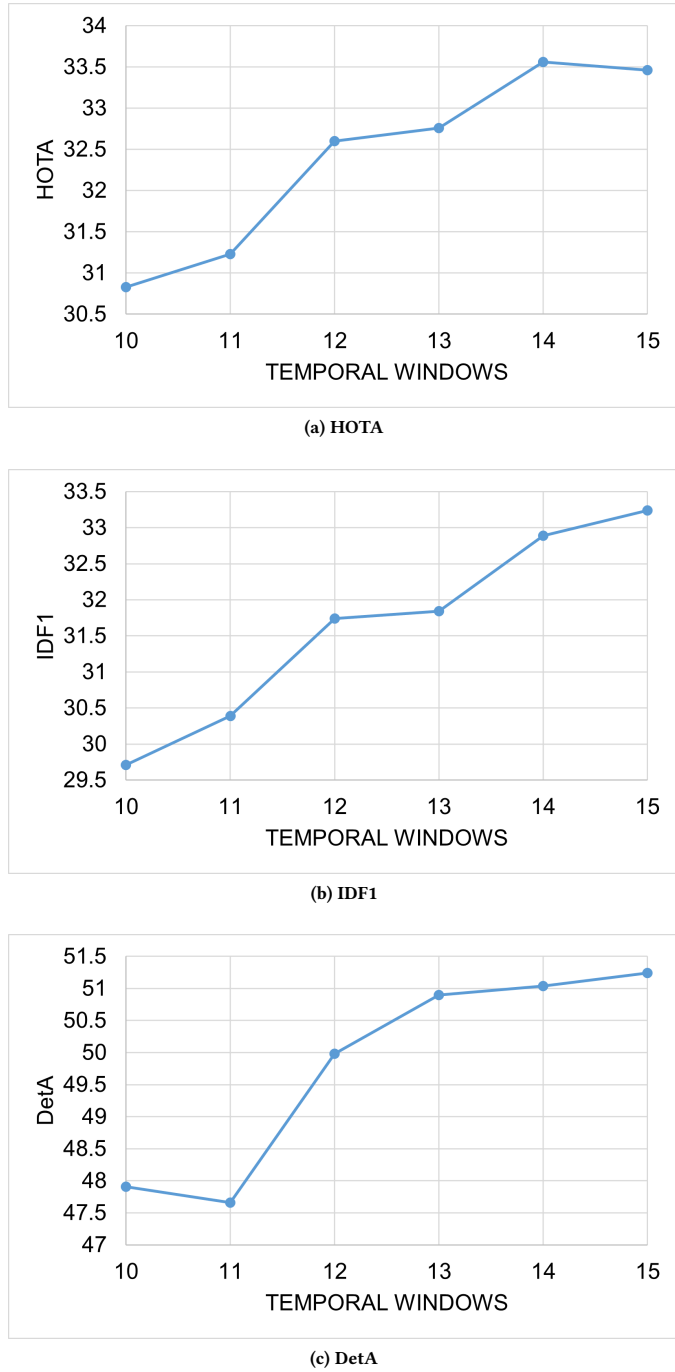


Figure C1: Comparison of different temporal windows.

the system. In this section, we examine the implications of varying time window sizes and their impact on the quality of tracking results.

### 3.1 Optimal Temporal Window Size

The temporal window size is paramount for accurately capturing sufficient temporal context to track objects, especially occlusions and rapid movements. Our experiments suggest that a time window that is too narrow may lead to insufficient data for reliable tracking. In contrast, an excessively wide window could introduce irrelevant information, increasing the computational burden and the potential for erroneous associations.

## 4 DEMO VISUALIZATION

The Demo Visualization section aims to elucidate the efficacy of our Ego3DT framework by presenting tangible examples of tracking and the influence of memory mechanisms. Through these visual demonstrations, we underscore the practical applications of our methodology and its performance in real-world scenarios.

### 4.1 Visualization of Tracking Results

The subsection on Tracking Results provides a compelling visual narrative of the 2D tracking capabilities facilitated by our underlying 3D association technique. The visualizations focus on the clarity with which objects are tracked across 2D video frames, attributing the success to the robust 3D scene construction at the heart of the Ego3DT framework. The qualitative demonstrations here highlight the precision and reliability of object tracking in 2D videos supported by our advanced 3D contextual understanding.

From Figure D2a, we compare the visualization results of ByteTrack and our method (Ego3DT) on the Ego3DT-daily dataset at moments  $T=30, 34, 42, 50$ , and  $51$ . The images reveal that ByteTrack struggles to track the green trash bin on the left side of the frame when people are walking. In contrast, Ego3DT can stably track the green trash bin ( $ID=112$ ).

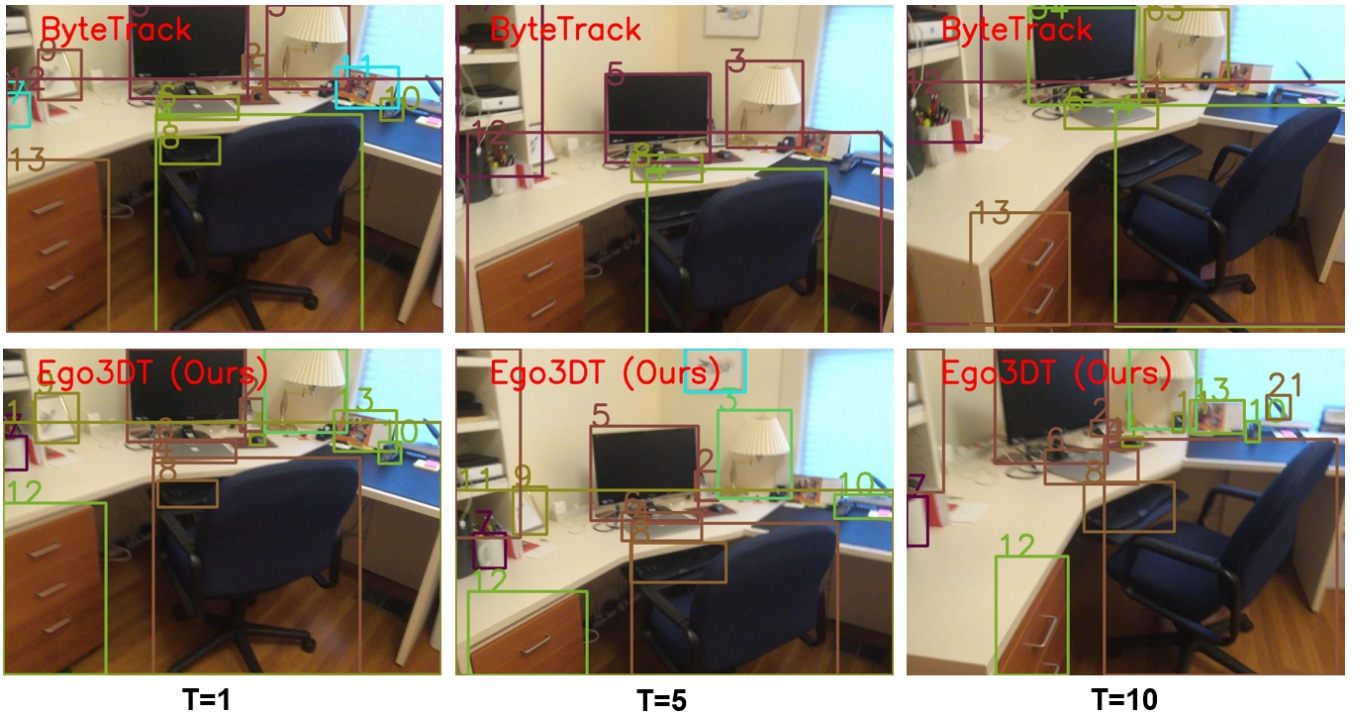
From Figure D2b, we compare the visualization results of ByteTrack and our method (Ego3DT) on the Ego3DT-indoor dataset at moments  $T=1, 5$ , and  $10$ . Due to the downsampling from 30 FPS to 3 FPS in Ego3DT-indoor, changes in camera view are quite pronounced. The images show that ByteTrack struggles to track the keyboard in the scene, and incorrectly assigns the keyboard's ID to the laptop ( $ID=8$ ). On the other hand, Ego3DT can stably track objects such as the monitor ( $ID=5$ ), laptop ( $ID=6$ ), teacup ( $ID=7$ ), keyboard ( $ID=8$ ), and photo ( $ID=13$ ). Therefore, this demonstrates that our method is more stable.

### 4.2 Impact of Memory

We conduct visualizations on the Ego3DT-indoor to demonstrate the role of the memory mechanism in 3D association. As shown in Figure D3c, we take the chair as an example and show the 3D points of the fifth frame matching with all



(a) Tracking results on Ego3DT-daily.



(b) Tracking results on Ego3DT-indoor.

**Figure D2: Visualization of Tracking Results on Ego3DT-daily and Ego3DT-indoor. The same global ID has the same color. (a) From T=30 to T=51, Ego3DT exhibits better tracking performance than ByteTrack, as demonstrated by the stable tracking of the green trash bin in the bottom left corner. (b) In indoor scenes of Ego3DT-indoor, Ego3DT also shows more stable tracking of objects, including monitors, laptops on desktops, and others, all maintaining consistent global IDs.**

the 3D points from the previous four frames. It can be seen that this method is more accurate compared to Figure D3b, which matches the 3D points of only one frame, resulting in a greater number of matched 3D points. This demonstrates

that the memory mechanism can effectively retain the 3D coordinates of objects, thereby helping to improve the stability of object tracking.



(a) Input sequence.



(b) The fifth frame matches the fourth frame.



(c) The fifth frame matches the first four frames.



(d) The reconstruction of 3D Field from input sequence.

**Figure D3: Visualization of Memory Mechanism on Ego3DT-indoor:** (a) The input sequence of frames  $T=1$  to  $T=5$  displays the tracking scenario. (b) The fifth frame's matching process with only the fourth frame, showing limited temporal context, is circled in red. (c) An enhanced matching approach in which the fifth frame matches the first four frames demonstrates the extended memory's role in capturing a broader temporal context for more accurate tracking. (d) A 3D field visualization further illustrates the depth and complexity of the scene, contextualizing the memory mechanism's impact on the tracking process.

REFERENCES

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. arXiv:2401.17270 [cs.CV]

[2] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, and Fisher Yu. 2022. Tracking Every Thing in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[3] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. 2023. OVTrack: Open-Vocabulary Multiple Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5567–5577.

[4] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.

[5] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. 2023. General object foundation model for images and videos at scale. *arXiv preprint arXiv:2312.09158* (2023).

[6] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*. Springer, 1–21.