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

Anonymous Authors

ABSTRACT

The growing interest in embodied intelligence has brought ego-centric perspectives to contemporary research. One significant challenge within this realm is the accurate localization and tracking of objects in ego-centric videos, primarily due to the substantial variability in viewing angles. Addressing this issue, this paper introduces a novel zero-shot approach for the 3D reconstruction and tracking of all objects from the ego-centric video. We present Ego3DT, a novel framework that initially identifies and extracts detection and segmentation information of objects within the ego environment. Utilizing information from adjacent video frames, Ego3DT dynamically constructs a 3D scene of the ego view using a pre-trained 3D scene reconstruction model. Additionally, we have innovated a dynamic hierarchical association mechanism for creating stable 3D tracking trajectories of objects in ego-centric videos. Moreover, the efficacy of our approach is corroborated by extensive experiments on two newly compiled datasets, with 1.04× - 2.90× in HOTA, showcasing the robustness and accuracy of our method in diverse ego-centric scenarios.

CCS CONCEPTS

• **Computing methodologies** → *Scene understanding*.

KEYWORDS

3D Vision, Open Vocabulary Tracking, Ego-centric Video

Figure 1: An illustrative example of Ego3DT. It showcases robust 3D object tracking across ego-centric video frames (from Frame 1 to Frame 5). The 3D field maintains consistent object information, ensuring the tracking ID remains unchanged. This delivers reliable tracking results in dynamic video scenarios, as shown by the persistent tracking of ID 1 and ID 2 across different viewpoints.

1 INTRODUCTION

Ego-centric, or first-person, computer vision addresses the perceptual challenges an embodied AI encounters in real-world situations. This area of research has garnered significant interest due to its relevance in various applications, including robotics [10, 55], embodied agents [70, 86–88], and augmented as well as mixed reality [16, 17, 35, 59]. One of the central tasks in this domain is multi-object tracking (MOT), which plays a critical role in numerous ego-centric applications. These applications range from monitoring the progress of actions or activities, re-identifying objects in one's environment, and forecasting the future states of the surrounding world.

Despite significant advancements in MOT, applying these methods to ego-centric videos remains underexplored. This gap is largely attributed to the absence of comprehensive ego-centric tracking datasets, essential for training and evaluating tracking algorithms [13]. Although the research community has introduced several popular tracking datasets such as OTB [74], TrackingNet [44], GOT-10k [21], and LaSOT [13], the high performance achieved by state-of-the-art trackers on these benchmarks does not effectively translate to ego-centric videos. This discrepancy underscores the urgent need for a dedicated ego-centric tracking dataset, particularly one that can support the unique requirements of ego-centric applications.

The distinct characteristics of ego-centric videos, as opposed to conventional third-person videos, pose unique challenges. These videos often capture a wide range of activities, objects, and locations without specific focus, reflecting the wearer's shifts in attention. Large head movements from the camera wearer frequently cause objects to exit and re-enter the field of view, and objects manipulated by hands may undergo frequent occlusions, along with rapid changes in scale, pose, and even state or appearance [57]. These unique aspects make object tracking significantly more demanding than in scenarios typically presented in existing datasets, highlighting a critical gap in current evaluation methodologies. Traditional MOT tasks [83], when applied to ego-centric videos [61], often result in poor tracking accuracy and robustness.

To better suit the variable conditions of ego-centric videos, our approach utilizes a 3D field representation, which offers a more adaptable and comprehensive framework for tracking. As shown in Figure 1, the 3D field captures the spatial layout and the temporal dynamics of objects within the scene, making it exceptionally suitable for the complexities of ego-centric views. This concept involves maintaining a dynamic 3D scene to enhance perceptual tasks [69, 70]. 3D perception can improve task robustness by ensuring stable object properties and relationships throughout the scene. By maintaining a dynamic 3D field, our approach preserves stable relationships and properties of 3D objects, significantly enhancing performance. Moreover, our method employs training-free, plug-and-play modules that deliver few-shot capabilities, distinguishing it from conventional approaches.

We summarize our contributions as follows:

- We propose a method for constructing a 3D scene from an ego-centric video and achieving open vocabulary object tracking, which requires only RGB videos as input and is a zero-shot approach.
- We implement object 3D position matching through a dynamic cross-window matching method, thereby alleviating the instability caused by relying solely on 2D image tracking.
- We achieve state-of-the-art performance on the open vocabulary multi-object tracking in ego-centric videos of daily activities, with 1.04× - 2.90× in HOTA.

2 RELATED WORK

2.1 Open Vocabulary Detection

Open vocabulary (OV) detection [79] has emerged as a novel approach to modern object detection, which aims to identify objects beyond the predefined categories. Early studies [18] followed the standard OV Detection setting [79] by training detectors on the base classes and evaluating the novel or unknown classes. However, this open-vocabulary setting, while capable of evaluating the detectors' ability to detect and recognize novel objects, is still limited to open scenarios and lacks generalization ability to other domains due to training on a limited dataset and vocabulary. Inspired by vision-language pre-training [23, 52], recent works [9, 26, 73, 89, 91] formulate open-vocabulary object detection as image-text matching and exploit large-scale image-text data to increase the vocabulary at scale. GLIP [32] presents a pre-training framework for open-vocabulary detection based on phrase grounding and evaluates in a zero-shot setting. Grounding DINO [39] incorporates the grounded pre-training into detection transformers [80] with cross-modality fusions. Several methods [36, 77, 78, 81] unify detection datasets and image-text datasets through region-text matching and pre-train detectors with large-scale image-text pairs, achieving promising performance and generalization. However, these methods often use heavy detectors, leading to high computational demands and deployment challenges. Utilizing the YOLO framework with an effective pretraining strategy, some works [4, 75] enhance open-vocabulary performance and generalization. GLEE [72] excels in recognizing and tracking objects across both images and videos.

2.2 Ego-centric Tracking

Over the last few decades, the introduction of numerous ego-centric video datasets [6, 14, 16, 28, 50, 60], has presented a wide range of fascinating challenges. Although many methodologies utilize tracking to address these challenges [7, 16, 29, 37, 41], it's notable that only a few studies have focused solely on the crucial issue of tracking. The works by Dunnhofer et al. [11, 12] address the specific challenges associated with ego-centric object tracking and represent the research most closely related to our own. However, a significant distinction exists in the scale of the dataset they utilized, which comprises 150 tracks designed purely for assessment purposes. In the realm of ego-centric video comprehension, Ego4D [16], EPIC-KITCHENS VISOR [7] and EgoTracks [61] are critical to our work. Ego4D stands out for its extensive compilation of ego-centric videos captured in natural settings and introduces numerous innovative tasks, including Episodic Memory, where tracking plays a pivotal role. Concurrently introduced, VISOR focuses on annotating brief videos (averaging 12 seconds in length) from EPIC-KITCHENS [6] with instance segmentation masks, illustrating the dynamic and detailed nature of research in this field.

2.3 Ego-centric 3D Understanding

The study of 3D object detection has made considerable advancements through the utilization of images [1, 20, 43, 54], point clouds [15, 30, 51, 58], and videos [2, 22]. To convert 2D images into 3D scenes, researchers have extensively employed Structure from Motion (SfM) techniques [47]. These techniques are divided into geometric-based methods [27, 45, 56], which rely on multiview geometry; learning-based methods [24, 67, 90], which utilize deep neural networks; and hybrid SfM approaches [62, 63], which integrate both strategies. SfM has been adapted for extensive videos in dynamic settings [85] and casual videos capturing everyday life [38, 84]. Yet, the distinct nature of ego-centric videos, characterized by their dynamic content, motion blur, and unconventional viewpoints, poses substantial hurdles to 3D scene understanding. While numerous studies have explored the reconstruction of 3D human poses from ego-centric footage [5, 19, 31, 53, 64, 68, 82], the field of 3D perception from an ego-centric perspective has seen exciting developments recently. HuCenLife [76] dataset is for large-scale human-centric scenarios, providing benchmarks for segmentation and action recognition tasks. Some work [48] tackled the challenges of ego-centric 3D human pose estimation with their domain-guided spatiotemporal transformer model, Ego-STAN, achieving significant improvements. EgoFish3D framework [40] employed self-supervised learning for accurate egocentric 3D pose estimation. Noteworthy efforts include investigation into ego-centric indoor localization using the Manhattan world assumption for room layouts [3], the development of EGO-SLAM for outdoor ego-centric videos through SfM over time [49], and the creation of NeuralDiff [66] and N3F [65], which innovate in dynamic NeRF technology for identifying and segmenting moving objects in ego-centric videos. Additionally, some work [46] proposed a method that correlates camera positions with video data to anticipate human-centric scene contexts. Our approach focuses on 3D tracking via dynamic matching in the 3D field. It is a zero-shot, RGB-only approach for open-vocabulary object tracking by constructing a 3D scene from an ego-centric video.

3 METHOD

3.1 Overview

As shown in Figure 2, **Ego3DT** is a purely vision-based open vocabulary 3D object tracking method $\mathcal{F}$ to achieve tracking results Y

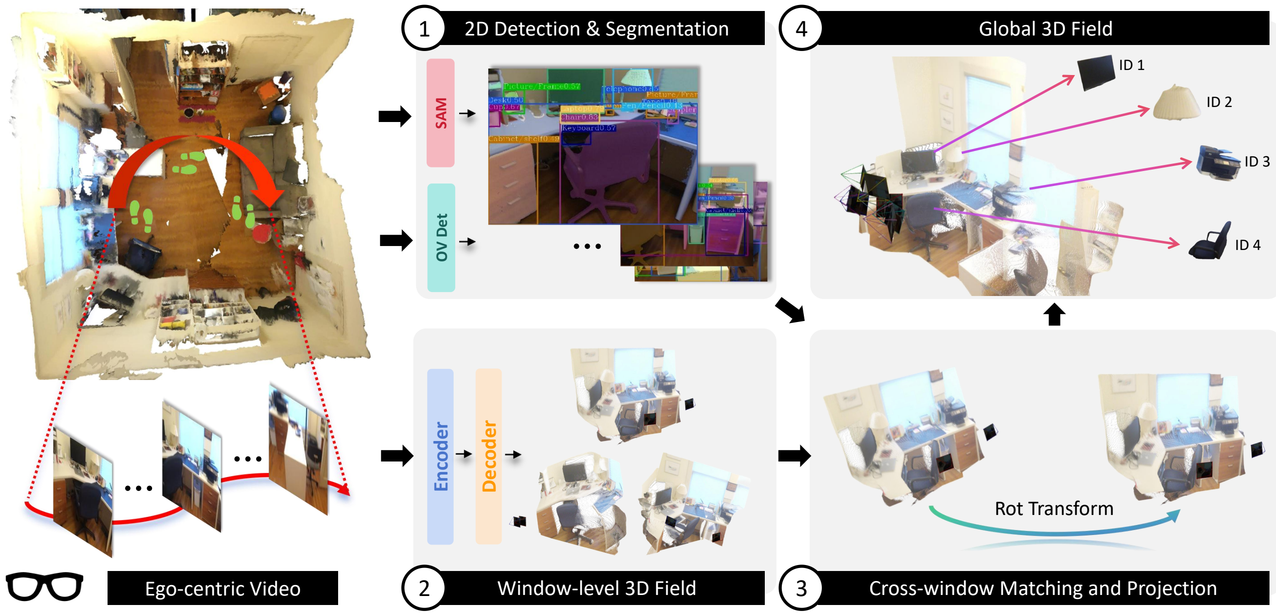

Figure 2: Ego3DT framework. (1) 2D Detection & Segmentation: Ego-centric video frames undergo object detection and segmentation using SAM to segment object points and an OV detector to identify objects. (2) Window-level 3D Field: The encoder-decoder structure processes the segmented frames to construct a window-level 3D field. (3) Cross-window Matching and Projection: Subsequent windows are aligned using rotational transforms to maintain object consistency across frames. (4) Global 3D Field: The cumulative data from all windows is integrated to form a global 3D field, with each object assigned a unique ID, facilitating precise object tracking throughout the video sequence.

from RGB ego videos X containing frames from I_1 to I_N. The open vocabulary object tracking results Y can be obtained as follows,

$$Y = \mathcal{F}(X), \quad X = [I_1, I_2, ..., I_N], \tag{1}$$

where $Y = \{O_i\}_{i \leq N}$ is the 3D object tracking output of the video with N frames, $O_i = [(x_j, y_j, z_j, \mathbf{ID}_j)]_{j \leq K}$ is a matrix containing 3D coordinates of tracked objects in each frame with identification **ID**, and K is the total number of tracked objects.

First, we conduct object detection **Det** on videos X, and semantic segmentation **Seg** based on detection output O_{2D}^{Det} as prompts:

$$O_{2D}^{Seg} = \mathbf{Seg}(O_{2D}^{Det}), \quad O_{2D}^{Det} = \mathbf{Det}(X), \tag{2}$$

where O_{2D}^{Seg} and O_{2D}^{Det} are the semantic segmentation and detection output respectively.

Then, we utilize a 3D estimation model $\mathcal{G}$ to map segmentation coordinates from 2D space O_{2D}^{Seg} to 3D space $O_{3D} \in \mathbb{R}^{K \times N \times 3}$:

$$O_{3D} = \mathcal{G}(X, O_{2D}^{Seg}), \tag{3}$$

where O_{3D} forms a one-to-one mapping between image pixels and 3D scene points, *i.e.*, $O_{2D} \leftrightarrow O_{3D}$, for all object coordinates $(x, y) \in \{1 \dots K\} \times \{1 \dots N\}$.

Finally, **Ego3DT** involves matching the 3D positions of objects using a hierarchical method, avoiding the instability issues that can arise from relying solely on 2D image tracking:

$$Y = \mathcal{M}(O_{3D}) = \mathbf{PointMatch}(\mathcal{A}(O_{3D})), \tag{4}$$

where the matching module $\mathcal{M}$ compares all the 3D points from frame to frame for precise object tracking Y with identification **ID**, and $\mathcal{A}$ is a 3D scene registration method aligning adjacent points. We use the additional Hungarian process to initialize matching **ID**.

3.2 2D Segmentation and Open-Vocab Detection

The foundational step in our method involves the precise identification and segmentation of objects within each frame of an ego-centric video. As shown in Equation (2), this process is bifurcated into two pivotal operations: 2D Open Vocabulary (OV) Detection **Det** and 2D Segmentation **Seg**, applied sequentially to the raw video frames to ensure a comprehensive understanding of the scene.

To achieve accurate object detection within our framework, we leverage the capabilities of the pretrained GLEE [72] in the experiment. Our efficient object detection model can identify a wide range of objects in 2D space across video frames, even those not explicitly labeled in the training data. We obtain precise 2D bounding boxes for all detectable objects by processing each frame through the model, setting the stage for subsequent segmentation.

Following the detection phase, the identified objects are further processed through SAM [25], a segmentation foundation model designed to delineate the precise boundaries of objects within an image. The bounding boxes obtained from GLEE [72] serve as prompts for SAM [25], enabling it to focus on specific regions of interest within the frame. This approach generates detailed segmentation maps for each object, including shape and location.

3.3 Window-level 3D Fields

We maintain window-level 3D fields with a 3D estimation model called $\mathcal{G}$, a pretrained DUSt3R [69]. This dual-branch system consists of image encoders, decoders, and regression heads. The image encoders are designed to extract detailed feature maps from segmented 2D object points, which are inputs derived from the preceding object detection and segmentation phases. The decoders then process these feature maps, which focus on extracting spatial relationships and depth cues from the encoded data. As shown in Equation (3), the 3D estimation model $\mathcal{G}$ processes segmented 2D object points, transforming them into their 3D counterparts through the pretrained DUSt3R [69]. This process is predicated on accurately detecting and segmenting objects within the 2D space, followed by their elevation into the 3D domain.

From 2D Segmentation to 3D Localization. As shown in Figure 2, the image encoders are designed to extract detailed feature maps from segmented 2D object points, inputs derived from the preceding object detection and segmentation phases. The decoders then process these feature maps, which focus on extracting spatial relationships and depth cues from the encoded data.

Integration and Alignment of 3D Data. The output from $\mathcal{G}$ consists of accurate 3D coordinates inherently aligned with the original RGB video frames. This alignment is critical as it ensures that each 3D point precisely represents its corresponding 2D point and is correctly positioned within the global context of the video sequence. This meticulous alignment facilitates the seamless integration of 2D and 3D data, enhancing the robustness and accuracy of the subsequent object-tracking processes.

Keeping window-level 3D fields in our framework advances the field of 3D object tracking and sets a new standard for the accuracy and efficiency of converting 2D video data into actionable 3D information. The rigorous processing and alignment of data ensure that our model is highly effective in the challenging environment of ego-centric videos, paving the way for innovative applications in multiple domains.

3.4 Cross-window Matching and Projection

The Matching Module $\mathcal{M}$ is a crucial component of the **Ego3DT** framework for tracking 3D objects across the video sequences. As shown in Equation (4), the Matching Module $\mathcal{M}$ consists of point-matching algorithms and a sliding window mechanism to ensure accurate and robust object tracking, even in occlusion or rapid movements. To minimize errors in point matching, we retain mutual correspondences between two images. This is achieved by performing KDTree search [69] in the 3D pointmap space.

Sliding Window Mechanism. We adapt the sliding window mechanism in the matching module, defined by the window size W, ensuring an overlap size T to maintain temporal continuity between frames. This design choice allows for the efficient processing of video frames by dividing the extensive task of 3D object tracking into manageable segments, each containing W frames. The step distance $S = W - T$ dictates the window's movement across the video sequence, ensuring that every frame is analyzed while optimizing computational resources.

Initial Object Tracking. The process begins by establishing a baseline of object tracking within the first window. For each frame i, up to the window size W, the 3D coordinates of detected objects $O^i_{3D} = \{(x_j, y_j, z_j)\}_{j \leq K}$ are determined, where K represents the number of objects detected within a frame. Utilizing KDTree distance calculations between every two consecutive frames, we employ the Hungarian algorithm to match objects based on their spatial proximity, thus assigning a unique Identification Number **ID** to each object. The result, Y_0, comprising tracked objects with their respective **ID**s within the first window, is stored in a buffer $\mathcal{B}$ for subsequent processing.

Dynamic matching across windows. As shown in the Algorithm 1, the module employs a hierarchical object-tracking approach. As the window slides by step S, each new set of frames is processed based on the previous window's data. Specifically, we employ a 3D scene registration method $\mathcal{A}$, an optimized homography process to align the 3D points of objects between the current and previous windows, thus $O^t_{3D} = \mathcal{A}(O^{t-1}_{3D}, O^t_{3D})$ to keep the current windows O^t_{3D} into the same space of the previous O^{t-1}_{3D}. The homography process is shown as follows:

$$O^{t-1}_{3D} = \prod_{t=1}^{T} H^t O^t_{3D}, \tag{5}$$

where H^t is the homography matrix between the current points O^t_{3D} and the previous points O^{t-1}_{3D} of overlapped frames, the ground points of all current frames are unified into the previous space. To further refine the alignment process, $\mathcal{M}$ employs an optimization strategy that minimizes the Euclidean distance between matched points across the homography transformations:

$$H^t_* = \arg\min_{\mathbf{KT}} \frac{1}{A} \sum_{t=1}^{T} ||O^{t-1}_{3D} - H^t O^t_{3D}||_2, \tag{6}$$

where A is the total number of matching points, H^t is a 4×4 matrix with rotation matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ and translation matrix $\mathbf{T} \in \mathbb{R}^{3\times1}$. All parameters are random numbers in the $(0, 1)$ range during initialization.

By recalculating KDTree distances for the newly aligned 3D points and based on the applied Hungarian algorithm, **PointMatch** matches pixels of objects from frame to frame. Each object in the current window is then assigned the **ID** of its closest match from the previous window, thus extending the tracking sequence. This process is repeated for each window throughout the video, culminating in comprehensively tracking all objects across the sequence.

The Matching Module $\mathcal{M}$ of **Ego3DT** achieves high precision and robustness in 3D object tracking through these sophisticated algorithms and mechanisms. It provides a global 3D field as shown in Figure 3. This innovative approach ensures that **Ego3DT** can effectively handle the complexities of ego-centric video analysis, paving the way for advancements in interactive and immersive technologies. We summarize the matching process in Algorithm 1, which outlines the step-by-step procedures for achieving accurate and reliable tracking results.

Table 1: Comparison of Open Vocabulary MOT performance. 2D box and 3D point refer to association to 2D box and 3D point. "f" stands for feature association.

Tracker	Detector	Association	HOTA (↑)	IDF1 (↑)	DetA (↑)	MT (↑)	ML (↓)	Frag (↓)
ByteTrack [83]	YOLO-World [4]	2D box	19.14	18.77	17.11	23	78	775
	GLEE [72]	2D box	29.58	**31.28**	29.10	**30**	73	1217
DeepSort [71]	YOLO-World [4]	2D box + f	10.63	9.63	11.15	9	106	637
	GLEE [72]	2D box + f	15.91	15.79	18	9	90	710
OVTrack [34]	OVTrack [34]	2D box + f	15.40	15.15	12.9	6	123	816
TET [33]	TET [33]	2D box + f	13.94	13.34	11.41	5	134	583
Ego3DT (Ours)	OVTrack [34]	3D point	13.44	12.9	13.79	5	138	512
	TET [33]		12.4	11.62	13.24	5	134	**463**
	YOLO-World [4]		16.28	15.28	19.43	14	78	1196
	GLEE [72]		**30.83**	29.71	**47.91**	24	**49**	1217

Algorithm 1 Cross-window Matching Process $\mathcal{M}$

Input: Video frames $X = \{I_i\}_{i=1}^{N}$, Initial object 3D coordinates O_{3D}^{1}, Window size W, Overlap size T
Output: Tracked objects Y with 3D coordinates and IDs
Initialize: Step size $S = W - T$, Buffer $\mathcal{B} \leftarrow \emptyset$
$Y_0 \leftarrow Hungarian(\mathbf{PointMatch}(O_{3D}^{1}))$
Add Y_0 to $\mathcal{B}$
for $t = 1$ to T **do**
  $O_{3D}^{t} \leftarrow \mathcal{G}(X, \mathbf{Seg}(\mathbf{Det}(I_t)))$
  Align 3D scenes: $O_{3D}^{t} \leftarrow \mathcal{A}(O_{3D}^{t-1}, O_{3D}^{t})$
  $Y_t \leftarrow \mathbf{PointMatch}(O_{3D}^{t-1}, O_{3D}^{t})$
  Add Y_t with IDs to $\mathcal{B}$
end for
Convert buffer $\mathcal{B}$ to the output space Y
return Y

4 EXPERIMENT

This section evaluates the **Ego3DT** framework for 3D object tracking in ego-centric videos using the Ego3DT Benchmark. We form two datasets: Ego3DT-daily and Ego3DT-indoor, and advance metrics to evaluate tracking accuracy. We test the state-of-the-art detectors and compare their performance to baseline models, demonstrating the efficacy and robustness of the **Ego3DT** in handling the unique challenges of ego-centric video analysis. Through rigorous testing and validation, this section illustrates the robustness, precision, and scalability of the **Ego3DT** framework.

4.1 Ego3DT Benchmark

Since there is no existing 3D object tracking benchmark based on ego-centric videos, we build a new benchmark called Ego3DT Benchmark to evaluate the performance of our model.

4.1.1 Datasets Description. We collected and re-annotated two datasets, Ego3DT-daily and Ego3DT-indoor, from Ego4D [16] and EmbodiedScan [70]. These datasets include 2D detection boxes and daily object trajectories in indoor and outdoor scenes.

Ego3DT-daily. contains six indoor and outdoor scenes, from which we collected videos from EGO4D. Each video, sampled at 10 FPS, consists of 500 consecutive frames. There are two outdoor scenes and four indoor scenes. The video collection locations include supermarkets, gardens, corridors, and kitchens. These ego-centric videos feature noticeable shaking and diverse object changes.

Ego3DT-indoor. includes data from five indoor scenes. Based on the Embodied Scan dataset, we collected ego-centric videos following predefined camera trajectories. We collected about 100 frames per video at 3 FPS from five scenes.

4.1.2 Annotation and Metrics. Our annotation pipeline is semi-automatic. We annotated the same objects with detection boxes and a global ID in a single video. For the Ego3DT-daily dataset, we first used the existing open vocabulary detector GLEE to extract object detection boxes to save annotation time. We then calibrated and aligned each object's detection boxes and IDs frame by frame. For objects that disappeared and then reappeared, we assigned them a consistent global ID. For the Ego3DT-indoor dataset, since Embodied Scan provides 3D detection boxes for each object, we projected the 3D detection boxes onto the current frame based on the camera's pose in each frame, thus determining the object's 2D detection boxes and global ID.

We evaluate the performance of our method using HOTA [42] and the MOT Challenge [8] evaluation metrics, including MOTA, IDF1, MT, ML, Frag *etc.* MOTA is computed based on false positives, false negatives, and ID switches and primarily focuses on detection performance. IDF1 assesses the consistency of IDs and places more emphasis on association performance. HOTA explicitly balances the accuracy of detection, association, and localization.

Table 2: Ablation study with different detectors and memory mechanisms of varying strengths.

Setting		HOTA (↑)	IDF1 (↑)	DetA (↑)	MT (↑)	ML (↓)	Frag (↓)
Detector	YOLO-World [4]	16.28	15.28	19.43	14	78	**1196**
	GLEE [72]	**30.83**	**29.71**	**47.91**	**24**	**49**	1217
Memory	w/o Memory	29.13	28.68	44.56	21	**49**	**1216**
	30 Frames	**30.83**	**29.71**	**47.91**	**24**	**49**	1217
	Full Frames	27.6	28.54	38.6	18	109	1241

4.2 Experiment Setups

We conduct experiments on **Ego3DT** using different detectors for open vocabulary detection, namely GLEE [72] via GLEE-Plus backbone Swin-L and YOLO-World [4] via YOLO-Worldv2-X. We also use SAM [25] with ViT-H backbone for open vocabulary segmentation. Then, we utilize the 3D estimation model via DUSt3R [69] with DPT Head, ViT-L Encoder, and ViT-B Decoder. Note that our experiments are conducted using only a single RTX3090-24G.

4.3 Baselines

We critically evaluate the **Ego3DT** framework against established baselines: ByteTrack [83], DeepSort [71], OVTrack [34], and TET [33], each offering unique strengths in multi-object tracking (MOT) and providing a comprehensive context for benchmarking our model's performance.

ByteTrack [83] is a powerful multi-object tracking (MOT) system designed to associate almost every detection box, regardless of the score, to improve tracking consistency, especially in cases with occluded objects. It stands out due to its simplicity, efficiency, and robustness against occlusions and low-confidence detections. The system has been successfully applied to different tracking benchmarks, confirming its versatility and strength as a baseline model for MOT tasks.

DeepSort [71] is an effective MOT method in videos, enabling accurate identity retention over time, particularly in scenarios where objects are frequently occluded. This system is a go-to choice for practitioners seeking a balance between performance and computational efficiency. The system proved versatile and robust, excelling as a baseline model in various MOT tasks.

OVTrack [34] is an open-vocabulary MOT method, utilizing vision-language models for classification and association, applying knowledge distillation and data hallucination techniques for feature learning. The approach aims to be highly data-efficient and is tailored for large-scale tracking, focusing on using static images for training.

TET [33] is a large-scale MOT method. It critically examines the limitations of current MOT metrics and methods, which often assume near-perfect classification performance, a presumption rarely met in practice. TET performs associations using Class Exemplar Matching, showing notable improvements in challenging tracking.

4.4 Evaluation Results

As shown in Table 1, we have evaluated the open-vocabulary multi-object tracking performance using a comprehensive range of metrics from the MOT Challenge and HOTA. **Ego3DT** greatly outperforms well-established baselines with a unique approach to 3D point association. It has been assessed on additional performance indicators, thus enhancing the breadth of our evaluation.

Notably, the **Ego3DT** framework with the GLEE detector [72] achieves the highest HOTA score of 30.83 among all evaluated trackers, indicative of a well-balanced detection and association accuracy. It excels in DetA (Detection Accuracy) with a leading score of 47.91, demonstrating our framework's exceptional capability in precise object detection. Note that DetA is not the same across different methods, even if the same detector is used. This is because different methods adopt different association and post-processing strategies that may affect the detection results. Furthermore, **Ego3DT** maintains a competitive edge with a high number of Mostly Tracked (MT) targets and the fewest Mostly Lost (ML) targets among the automatic tracking methods, with respective scores of 24 and 49, highlighting the framework's robustness in persistent object tracking over time. The TET detector [33] produces the highest number of Fragmentations (Frag), indicating that our tracking is accurate and the object identity is stable compared to the real object trajectories in the ground truth data.

These expanded metrics provide a holistic view of our framework's performance, affirming its strengths in maintaining object identities (as evidenced by its IDF1 score of 29.71) and effectively tracking objects throughout the video sequence. Despite the high Frag count, the **Ego3DT** framework's overall leading performance in key metrics solidifies its status as a robust solution for the MOT challenge, particularly within the demanding context of ego-centric videos.

4.5 Ablation Study

To refine the **Ego3DT** framework, we conduct a comprehensive ablation study to discern the individual contributions of detector quality and memory mechanisms to the framework's overall performance. The experiments are carefully designed to isolate the impact of these components, providing insights into their respective significance and interplay. As shown in Table 2, a high-quality detector

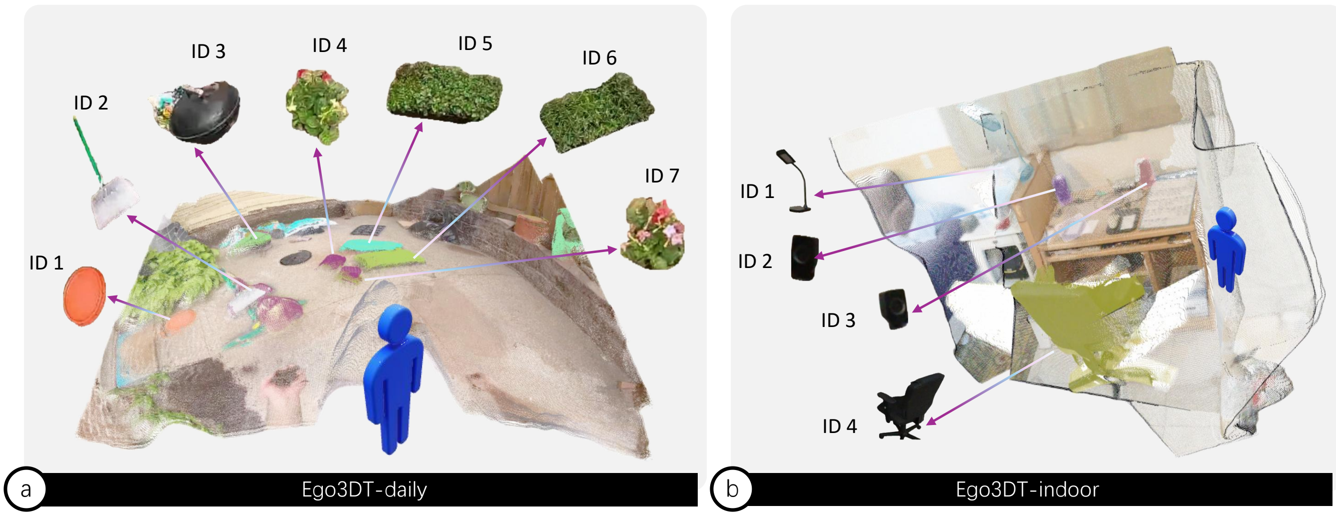

Figure 3: Qualitative results of the 3D tracking field in Ego3DT: a) For the Ego3DT-daily dataset, diverse outdoor objects (IDs 1-7) are successfully tracked within the environment, showing the model's capability to handle varying object types and outdoor conditions. b) In the Ego3DT-indoor dataset, common indoor objects (IDs 1-4) are tracked with high fidelity in a typical room setup, demonstrating the precision of the 3D tracking across different indoor scenes.

profoundly influences the framework's performance, and memory mechanisms play a nuanced role in achieving state-of-the-art tracking performance in open vocabulary MOT scenarios.

Accurate Detector is Pivotal. The choice of detector plays a pivotal role. The GLEE detector [72], being a high-quality pretrained detector, synergizes exceptionally well with our 3D association approach, yielding a HOTA score of 30.83, which robustly indicates superior detection and ID association. This confirms that a proficient detector, coupled with our sophisticated 3D tracking methodology, can substantially boost the overall tracking quality, ensuring that high-quality detections translate into high-quality ID annotations.

Appropriate Memory Mechanism is Critical. It reflects on the balance between memory usage and tracking performance. Notably, using a 30-frame memory mechanism offers the best performance across all metrics. This optimized setting achieves a HOTA of 30.83 and an IDF1 of 29.71, underscoring the effectiveness of a limited temporal memory that captures the immediate past to maintain context without being burdened by the noise of distant frames. On the other hand, the absence of a memory mechanism and the use of full frame memory result in reduced performance, demonstrating the importance of a focused temporal window for accurate tracking. This suggests that an excessive memory span can dilute the relevancy of information, leading to higher fragmentation and decreased detection accuracy. The results highlight the delicate trade-off between the memory's depth and the tracking accuracy, suggesting that moderate memory size is instrumental in improving the consistency and precision of object tracking in ego-centric videos.

4.6 Qualitative Results

Our **Ego3DT** framework exhibits significant advancements in 3D reconstruction and 2D tracking, showcasing robust performance even under challenging first-person motion scenarios. We provide a qualitative analysis of these two core aspects to highlight the efficacy and improvements over existing methodologies.

4.6.1 Qualitive Results on 3D Reconstruction. A closer examination of **Ego3DT**'s performance on our meticulously collected datasets reveals the nuanced capability of our framework in handling complex 3D environments. As shown in Figure 3, we present two distinct scenarios that showcase the efficacy of **Ego3DT** in real-world applications.

Outdoor Tracking in Ego3DT-daily. The Ego3DT-daily dataset, representing an array of outdoor settings, challenges the framework with dynamic lighting, diverse object shapes, and sizes. Our model demonstrates robustness in these conditions, accurately tracking and maintaining consistent IDs across different object types, from smaller items like a wok (ID 3) to larger potted plants (IDs 5 and 6). The 3D tracking field captures the spatial relations and movement paths, illustrating the model's adaptability to outdoor environments.

Indoor Persistence in Ego3DT-indoor. Transitioning to the indoor domain, the Ego3DT-indoor dataset offers a contrasting setting with more controlled lighting but equally complex object interactions. The model successfully delineates and tracks objects such as furniture (IDs 1 to 4) in a typical room scenario, highlighting its precision in cluttered, confined spaces. The tracking continuity is evident, with the framework skillfully handling occlusions and varying distances from the camera.

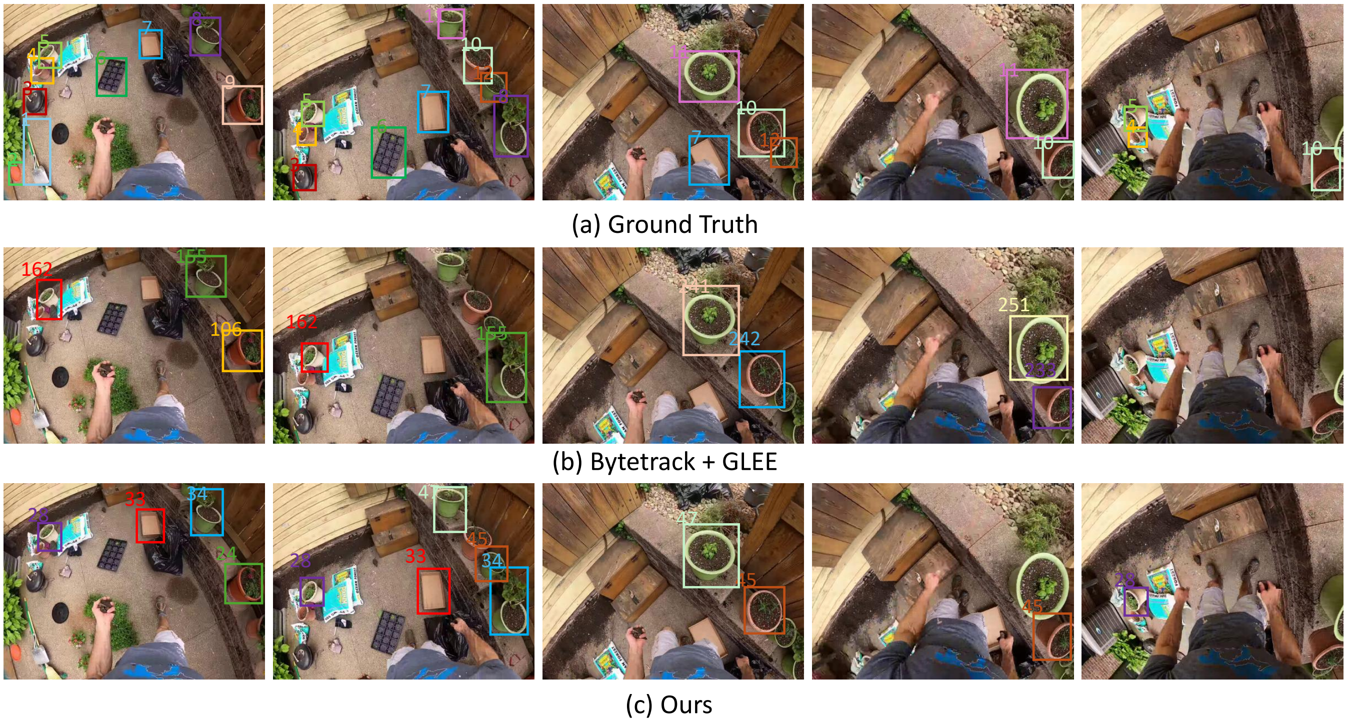

Figure 4: Qualitative results of 2D tracking comparison: a) Ground Truth sequence showing accurate object detection and consistent ID assignment over time. b) ByteTrack with GLEE detection demonstrating object tracking and identification, with occasional ID inconsistencies and missed detections. c) Our Ego3DT approach, which maintains stable object identification, accurately captures dynamic objects and shows superior consistency in ID assignment, especially evident in motion-rich ego-centric perspectives. From left to right, the frames progress temporally, illustrating each method's tracking continuity and precision.

4.6.2 Qualitive Results on 2D Tracking. As shown in Figure 4, we compare our method to ByteTrack [83] and demonstrate how our approach offers improved detection stability. Our framework, **Ego3DT**, is particularly effective in dynamic scenes that involve uniform motion from a first-person perspective. It consistently identifies and tracks a higher number of objects with greater reliability, making it an ideal solution for applications that require real-time responsiveness and accuracy, such as augmented reality and autonomous navigation systems.

As seen in the side-by-side comparison of 2D tracking techniques, our **Ego3DT** framework excels at preserving object identity across frames. The Ground Truth (a) provides a benchmark with flawless tracking and ID fidelity. ByteTrack coupled with GLEE detection (b) provides a strong baseline but occasionally falters with ID switches and detection lapses, especially under the erratic motion intrinsic to ego-centric videos. In contrast, our approach (c) consistently demonstrates a remarkable grasp on object trajectories, maintaining accurate IDs even in the presence of motion blur and rapid scene changes. Moving from left to right through the temporal sequence, this comparison underscores the Ego3DT framework's advanced capability to deliver reliable and coherent tracking performance in dynamic and challenging first-person video scenarios.

These qualitative results underscore the versatile 3D tracking capabilities of **Ego3DT**, cementing its potential for comprehensive scene understanding and robust object tracking in diverse environments.

5 LIMITATION AND FUTURE WORK

Although our proposed **Ego3DT** can successfully detect and track almost every 3D object in the scene, it might still fail in tracking some rapidly moving objects like cats, dogs, or humans. We leave this in future work, including tracking the moving objects and detecting the interaction with the scene and other objects.

6 CONCLUSION

We have introduced the **Ego3DT** framework for accurately tracking 3D objects in ego-centric videos. The framework uses a sophisticated 3D estimation model and state-of-the-art detection and segmentation technologies. Our experimental results demonstrate that the **Ego3DT** framework outperforms established baselines and can ensure accurate detection, consistent ID tracking, and precise localization. The **Ego3DT** framework can facilitate practical applications in augmented reality, robotics, and advanced surveillance systems.

REFERENCES

[1] Wentao Bao, Bin Xu, and Zhenzhong Chen. 2019. Monofenet: Monocular 3d object detection with feature enhancement networks. *IEEE Transactions on Image Processing* 29 (2019), 2753–2765.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[3] Xiaowei Chen and Guoliang Fan. 2022. Egocentric Indoor Localization From Coplanar Two-Line Room Layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1549–1559.

[4] 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]

[5] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. 2022. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6792–6802.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*.

[7] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. 2022. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[8] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. 2021. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision* 129 (2021), 845–881.

[9] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *CVPR*. 14064–14073.

[10] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2022).

[11] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. 2021. Is first person vision challenging for object tracking?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2698–2710.

[12] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. 2023. Visual Object Tracking in First Person Vision. *International Journal of Computer Vision* 131, 1 (2023), 259–283.

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 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.

[14] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1226–1233. https://doi.org/10.1109/CVPR.2012.6247805

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2023. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259* (2023).

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*.

[19] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. 2021. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4318–4329.

[20] Abdullah Hamdi, Bernard Ghanem, and Matthias Nießner. 2022. SPARF: Large-Scale Learning of 3D Sparse Radiance Fields from Few Input Images. *arXiv preprint arXiv:2212.09100* (2022).

[21] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 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 (2019), 1562–1577.

[22] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. 2018. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 954–960.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*. 4904–4916.

[24] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*. 2938–2946.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[26] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. 2022. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. *CoRR* abs/2209.15639 (2022). arXiv:2209.15639

[27] Mathieu Labbé and François Michaud. 2019. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics* 36, 2 (2019), 416–446.

[28] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1346–1353. https://doi.org/10.1109/CVPR.2012.6247820

[29] Yong Jae Lee and Kristen Grauman. 2015. Predicting Important Objects for Egocentric Video Summarization. *Int. J. Comput. Vision* 114, 1 (aug 2015), 38–55. https://doi.org/10.1007/s11263-014-0794-5

[30] Bing Li, Cheng Zheng, Silvio Giancola, and Bernard Ghanem. 2022. SCTN: Sparse Convolution-Transformer Network for Scene Flow Estimation. In *AAAI*.

[31] Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17142–17151.

[32] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded Language-Image Pre-training. In *CVPR*. 10955–10965.

[33] 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)*.

[34] 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.

[35] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. 2021. Ego-Exo: Transferring Visual Representations from Third-person to First-person Videos. In *CVPR*.

[36] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. 2023. Learning Object-Language Alignments for Open-Vocabulary Object Detection. In *ICLR*.

[37] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. 2019. Forecasting Human Object Interaction: Joint Prediction of Motor Attention and Egocentric Activity. *ArXiv* abs/1911.10967 (2019).

[38] Sheng Liu, Xiaohan Nie, and Raffay Hamid. 2022. Depth-Guided Sparse Structure-from-Motion for Movies and TV Shows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15980–15989.

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *CoRR* abs/2303.05499 (2023). arXiv:2303.05499 https://doi.org/10.48550/arXiv.2303.05499

[40] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. 2023. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia* (2023).

[41] Cewu Lu, Renjie Liao, and Jiaya Jia. 2015. Personal object discovery in first-person videos. *IEEE Transactions on Image Processing* 24, 12 (2015), 5789–5799. https://doi.org/10.1109/TIP.2015.2487868

[42] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision* 129 (2021), 548–578.

[43] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. 2021. M3dssd: Monocular 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6145–6154.

[44] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 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.

[45] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics* 31, 5 (2015), 1147–1163.

[46] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. 2022. Egocentric scene context for human-centric environment understanding from video. *arXiv preprint arXiv:2207.11365* (2022).

[47] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. 2017. A survey of structure from motion*. *Acta Numerica* 26 (2017), 305–364.

[48] Jinman Park, Kimathi Kaai, Saad Hossain, Norikatsu Sumi, Sirisha Rambhatla, and Paul Fieguth. 2023. Domain-Guided Spatio-Temporal Self-Attention for Egocentric 3D Pose Estimation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1837–1849.

[49] Suvam Patra, Kartikeya Gupta, Faran Ahmad, Chetan Arora, and Subhashis Banerjee. 2019. Ego-slam: A robust monocular slam for egocentric videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 31–40.

[50] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 2847–2854.

[51] Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. 2022. Pix4point: Image pretrained transformers for 3d point cloud understanding. *arXiv preprint arXiv:2208.12259* (2022).

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.

[53] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–11.

[54] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. 2022. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2397–2406.

[55] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9339–9347.

[56] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. 2020. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9869–9878.

[58] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.

[59] Maximilian Speicher, Brian D Hall, and Michael Nebeling. 2019. What is mixed reality?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[60] Yu-Chuan Su and Kristen Grauman. 2016. Detecting Engagement in Egocentric Video. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 454–471.

[61] Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. 2024. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems* 36 (2024).

[62] Zachary Teed and Jia Deng. 2018. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605* (2018).

[63] Zachary Teed and Jia Deng. 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems* 34 (2021), 16558–16569.

[64] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. 2020. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *arXiv preprint arXiv:2011.01519* (2020).

[65] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. 2022. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494* (2022).

[66] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. 2021. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 910–919.

[67] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. 2017. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804* (2017).

[68] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. 2021. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11500–11509.

[69] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2023. DUSt3R: Geometric 3D Vision Made Easy. *arXiv preprint arXiv:2312.14132* (2023).

[70] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. 2024. EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[71] 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.

[72] 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).

[73] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023. Aligning Bag of Regions for Open-Vocabulary Object Detection. In *CVPR*. 15254–15264.

[74] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2411–2418.

[75] Johnathan Xie and Shuai Zheng. 2021. ZSD-YOLO: Zero-Shot YOLO Detection using Vision-Language KnowledgeDistillation. *CoRR* (2021). arXiv:2109.12066

[76] Yiteng Xu, Peishan Cong, Yichen Yao, Runnan Chen, Yuenan Hou, Xinge Zhu, Xuming He, Jingyi Yu, and Yuexin Ma. 2023. Human-centric scene understanding for 3d large-scale scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20349–20359.

[77] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 2023. DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-training via Word-Region Alignment. In *CVPR*. 23497–23506.

[78] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. 2022. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection. In *NeurIPS*.

[79] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-Vocabulary Object Detection Using Captions. In *CVPR*. 14393–14402.

[80] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2023. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *ICLR*.

[81] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. GLIPv2: Unifying Localization and Vision-Language Understanding. In *NeurIPS*.

[82] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*. Springer, 180–200.

[83] 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.

[84] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. 2022. Structure and motion from casual videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 20–37.

[85] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. 2022. ParticleSfM: Exploiting Dense Point Trajectories for Localizing Moving Cameras in the Wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 523–542.

[86] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Li Boyi, Shengyu Hao, Shidong Cao, Tian Ye, Jenq-Neng Hwang, and Gaoang Wang. 2023. See and think: Embodied agent in virtual environment. *arXiv preprint arXiv:2311.15209* (2023).

[87] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. 2024. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282* (2024).

[88] Zhonghan Zhao, Ke Ma, Wenhao Chai, Xuan Wang, Kewei Chen, Dongxu Guo, Yanting Zhang, Hongwei Wang, and Gaoang Wang. 2024. Do We Really Need a Complex Agent System? Distill Embodied Agent into a Single Model. *arXiv preprint arXiv:2404.04619* (2024).

[89] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2022. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*. 16772–16782.

[90] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1851–1858.

[91] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In *ECCV*. 350–368.