Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind

Chiara Plizzari[1]* Shubham Goel[2,3] Toby Perrett[4] Jacob Chalk[4]
Angjoo Kanazawa[3] Dima Damen[4]

[1]Politecnico di Torino [2]Avataar Inc. [3]University of California, Berkeley
[4]University of Bristol

http://dimadamen.github.io/OSNOM

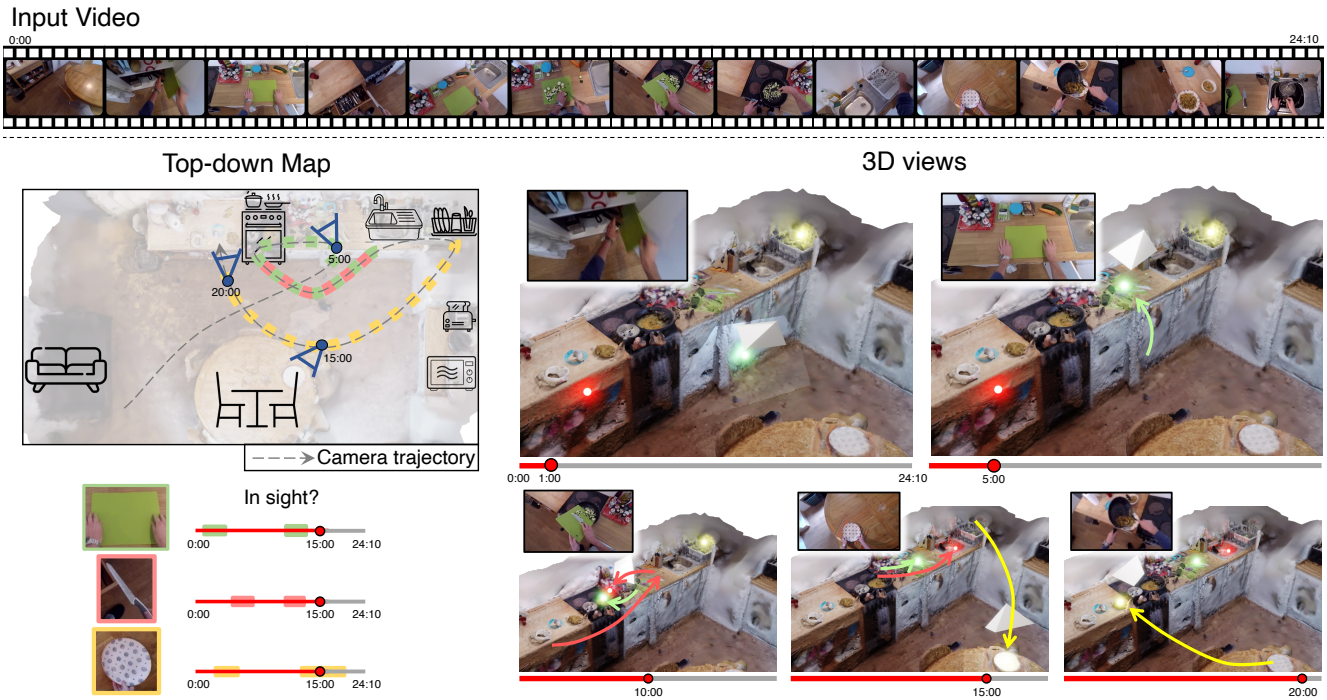

Figure 1. **Spatial Cognition.** From an egocentric video (top), we propose the task *Out of Sight, Not Out of Mind*, where the 3D locations of all active objects are known when they are both in- and out-of-sight. We show a 24 mins video along with world-coordinate tracks of 3 active objects through the video – from a top-view down with camera motion (left top); identifying when they are in-sight (left bottom); their trajectory from a side view at five different frames (right). Neon balls show the 3D locations of these objects over time along with the camera (white prism), corresponding frame (inset) and object location change (coloured arrow). The chopping board is picked from a lower cupboard (1:00) and is in-hand at 05:00. The knife is picked up from the drawer (after 05:00), while in use (10:00) until it is discarded in the sink (before 15:00). The plate travels from the drainer to the table (15:00), then back to the counter (20:00).

Abstract

As humans move around, performing their daily tasks, they are able to recall where they have positioned objects in their environment, even if these objects are currently out of their sight. In this paper, we aim to mimic this spatial cognition ability. We thus formulate the task of Out of Sight, Not Out of Mind – 3D tracking active objects using observations captured through an egocentric camera. We introduce a simple but effective approach to address this challenging problem, called Lift, Match, and Keep (LMK). LMK ***lifts*** *partial 2D observations to 3D world coordinates,* ***matches*** *them over time using visual appearance, 3D location and interactions to form object tracks, and* ***keeps*** *these object tracks even when they go out-of-view of the camera. We benchmark LMK on 100 long videos from EPIC-KITCHENS. Our results demonstrate that spatial cognition is critical for correctly locating objects over short and long*

*Work carried during Chiara's research visit to the University of Bristol

time scales. E.g., for one long egocentric video, we estimate the 3D location of 50 active objects. After 120 seconds, 57% of the objects are correctly localised by LMK, compared to just 33% by a recent 3D method for egocentric videos and 17% by a general 2D tracking method.

1. Introduction

It is lunch time and the pan is on the hob. You bend to pick the chopping board from a lower cupboard and put it on the counter. You then retrieve a knife from the cutlery drawer. You use the chopping board and knife to slide chopped food into the pan before discarding both in the sink. You then retrieve a clean plate from the drainer to serve the food. As you move around the kitchen, you are aware of where these objects are even if they are currently out of view.

This ability to "know what is where" is an integral part of *spatial cognition*. It allows humans to build a mental map of their environment, including "memories of objects, once perceived as we moved about" [11]. Importantly, spatial cognition dictates these objects exist independently of human attention, and continue to exist in the cognitive map when the observer has left the vicinity [4, 6, 28, 58]. Spatial cognition is an innate ability, crucial to human survival, as it is how humans "acquire and use knowledge about their environment to determine where they are, how to obtain resources, and how to find their way home" [50].

In this paper, we operate on egocentric videos and make three prime **[C]**ontributions.

[C1] We introduce the task Out of Sight, Not Out of Mind (OSNOM) – maintaining the knowledge of where *all objects* are, as they are moved about and even when absent from the egocentric video stream. Egocentric views allow detailed observation of objects during interactions, e.g. the camera can look into the fridge or oven, and see exactly what was picked from the drainer. However, objects often swiftly move out of the camera's field of view. We focus on these challenging set of *active objects* that are moved by the camera wearer during the video sequence. Our task is to position **multiple dynamic** objects in 3D throughout the video, **both in- and out-of-view**. This is distinct from existing tasks, such as episodic memory [16], which search for the presence of an object within the video, i.e. within the camera's field of view. Instead, the OSNOM task evaluates the locations of objects even when they are out of sight. Figure 1 illustrates the OSNOM task.

To address the OSNOM challenge, **[C2]** we propose a simple but effective approach that tracks objects in the world coordinate frame. Specifically, we *lift* observations to 3D – by reconstructing the scene mesh and projecting 2D detections given their depth from camera and surface estimates. We then *match* these lifted observations using appearance and location over time to form consistent object tracks, and *keep* the knowledge of objects in mind when they are out of sight. This *lift, match and keep (LMK)* approach allows spatio-temporal understanding which humans take for granted, yet is out of reach of current methods — knowing when an object is within reach but is out-of-view, or when in-view but occluded inside a cupboard.

[C3] We benchmark OSNOM on 100 long videos from the EPIC-KITCHENS dataset [7] through past and future 3D location estimations over multiple timescales. We showcase that objects are out of view for 85% of frames on average. Using our LMK approach, we can correctly position 64% of the objects successfully after 1 minute, 48% after 5 minutes, and 37% after 10 minutes, consistently outperforming recent approaches for ego [26, 61] and general [60] tracking. Ablations demonstrate that maintaining 3D object locations over time is critical for correctly locating moving objects, and when they are occluded or out of view.

2. Related Works

Egocentric vision has traditionally focused on tasks within the recorded video stream, i.e. within the camera's field of view. These include understanding actions, objects and interactions over short, and more recently longer [7, 8, 16, 46], timescales. Even when addressing future prediction (*e.g.* action anticipation [14]), memory (*e.g.* episodic memory [16]), object tracking [46], approaches scan the video stream to find when an object is in-sight. The seminal work Ego-Topo [29] builds a 2D affordance graph of the environment, relating actions to automatically discovered hotspots. The motivation to capture the relative location of an object to the camera wearer was explored in EgoEnv [30], by pre-training on 3D simulated environments. It shows that such environmentally-informed representations can improve performance on down-stream tasks such as episodic memory.

A number of tasks have been recently proposed that require 3D understanding in egocentric vision, such as jointly recognising and localising actions in a 3D map [24]. In Visual Query localisation in 3D (VQ3D) [16], given a query image of an object, the aim is to localise *only one* 3D position – when *the object was last seen unoccluded and in view.* Tracking is thus unnecessary (*e.g.* SOTA on VQ3D, EgoLoc [26], is based on retrieval and notes most objects are stationary). Recently, Zhao *et al.* [61] propose tracking a single object in the world-coordinate frame from RGB-D videos. They track objects in 2D, then lift these to 3D using the sensor's depth. Another recent work [18] estimates camera pose from Dust3r [51], to predict positions of static objects like parts of the sofa. These are then tracked using rotational transforms to maintain 3D object consistency. Our approach is complementary in that we do not track objects that remain static but instead focus on tracking objects that the person has moved. Importantly, both approaches only track objects when in view. Distinctly, the OSNOM task locates objects both when in- and out-of-view, offering

a complete spatial cognition of dynamic objects.

3D egocentric datasets are now becoming available [8, 16, 17, 32, 37]. Examples include Ego4D [16], which provides 3D scans and sparse camera poses for 13% of the dataset, Ego-Exo4D [17] which captures multiple first- and third-person views, and the Aria Digital Twin [32], which contains both camera poses and object segmentation masks for its two environments. EPIC-Fields [48] provides a pipeline to extract point clouds and dense camera poses from egocentric videos, and provides camera estimates for the EPIC-KITCHENS dataset [7] across 45 kitchens. We use the pipeline from EPIC-Fields [48] paired with dense active object masks from VISOR [8].

Most related to 3D object tracking, [61] provides a small-scaled dataset with instance-level annotations in both 2D and 3D (4.5 hours, 250 objects in 10 environments). Instead, we use a more diverse dataset collected in an unscripted manner, with our annotations covering 25 hours, more than 2K objects in 45 environments. Similarly Ego3DT [18] is evaluated on very short videos (less than 1 minute long).

Object tracking through occlusion has been investigated in 2D, where maintaining object permanence, through heuristics [19] (*e.g.* constant velocity [3]) or learning [42, 47], can track assignment when occluded objects reappear. However, these works do not track out of the field of view, and evaluate on short-term sequences (*e.g.* TCOW [49] uses sequences of maximum length of 464 *frames*).

Autonomous-driving typically maintains a map of the vehicle's surroundings [52] and tracks nearby vehicles, even when out of sight. However, whilst maintaining object locations through occlusion [13, 38], tracks are deleted regularly as the vehicle only has to know about its vicinity.

Human tracking has seen progress from 2D [1, 27, 60], to 3D [35], to 3D with motion models [15, 21, 36] which predict the location of occluded humans. Although these approaches use 3D for tracking, they usually do so in the camera coordinate frame. Recent works have explored simultaneous reconstruction of camera motion and human pose in 3D [22, 55, 56], with [45, 55] evaluating this concept on human tracking. [20] proposes a benchmark for tracking humans from multiple ego- and exo-centric cameras.

We present the first egocentric video work to track multiple objects in the world coordinate frame. Unlike humans, objects in egocentric videos do not move autonomously and frequently enter and exit the camera view. Our approach focuses on dynamic objects, tracking them in 3D space even when out of view, preserving object permanence. We detail our approach next.

3. Method - Lift, Match and Keep (LMK)

Our method operates on a single untrimmed egocentric video, E, recorded in an indoor environment. We aim to keep track of all objects of interest in the 3D world coordinate frame. These 3D tracks capture the locations of objects throughout the video, even when they are not visible in the camera frame, solving the task of Out of Sight, Not Out of Mind (OSNOM).

As many objects in the scene remain in the same position throughout the video, we focus on the challenging set of active objects that the camera wearer interacts with, typically moving these objects from one place to another, often multiple times in the video.

We take as input observations of active objects $o_n = (f_n, m_n)$, where f_n is a frame, and m_n is a semantic-free 2D mask in that frame given in *image coordinates*. The set of all observations, across the whole video, is $\mathcal{O} = \{o_n : n = 1, ..., N\}$. We call these observations *partial*, as they do not exist for every object in every frame. The number of observations N is much larger than the number of active objects - each object may be the subject of multiple observations. N is also independent of the number of frames T, as frames may contain zero or multiple masks.

We call our method Lift, Match and Keep (LMK). We first *lift* 2D observations of objects to 3D (Sec 3.1), *match* them over time, and *keep* objects in mind when they are out-of-sight (Sec 3.2). We detail LMK next.

3.1. Lift: Lifting 2D Observations to 3D

3D Scene Representation. Given a single egocentric video stream, we follow the pipeline proposed in [48] to estimate camera poses and a sparse point cloud of the static scenes. We ignore redundant frames by calculating the homography over consecutive frames, thus allowing these long videos to be processed by Structure from Motion (SfM) pipelines such as COLMAP [40]. The selected subset of video frames contains sufficient visual overlap to register all frames to the SfM point cloud and estimate a camera pose C_t for every time t in the video. Note that the intrinsic parameters of the camera are also automatically estimated by this pipeline.

This reconstruction focuses on estimating the static background of the scene. Objects in motion are deemed as outliers during matching and are accordingly ignored in the reconstructions. The pipeline produces a sparse point cloud that cannot be used for positioning objects in 3D as it is missing the notion of surfaces. We convert these point clouds to surface representations as follows.

We extract scene geometry as a 3D mesh using a classical Multi-View Stereopsis pipeline [12, 41] that runs patch matching to find dense correspondences between stereo image pairs, triangulates the correspondences to estimate depth, and fuses them together into a dense 3D point cloud with surface normals. We recover a scene mesh $\mathcal{S}$ from the dense point cloud using Delaunay surface reconstruction [5]. Examples of these meshes can be seen in Figure 1.

Estimating 3D locations from monocular depth. For each

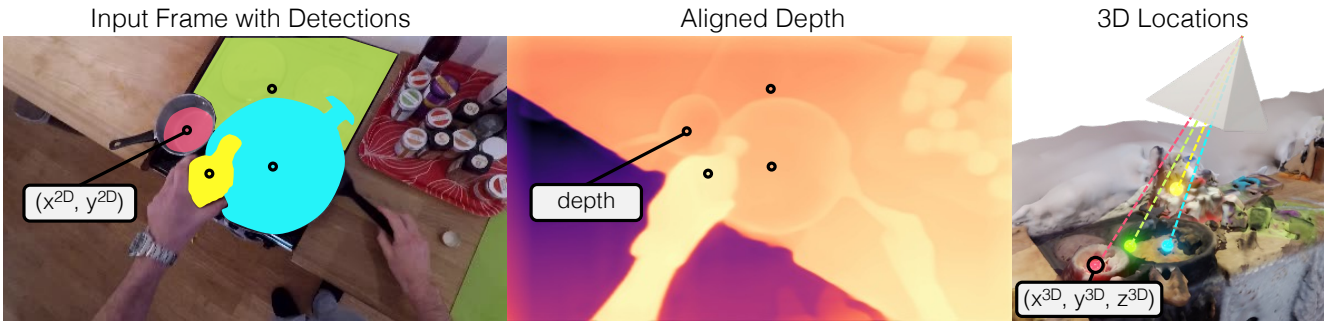

Figure 2. **Lifting 2D observations to 3D.** We use mask centroids as 2D object locations, sample corresponding depths from the mesh-aligned monocular depth estimate. We then compute the 3D object locations in world coordinates by un-projecting the mask's centroid from the estimated camera pose.

frame, f_n, we estimate the monocular depth estimation using [53]. The advantage of using this approach is the ability to estimate the position of both static and dynamic objects, including objects that are in-hand. However, this per-frame depth is incorrectly scaled and temporally inconsistent. We thus align it to the reconstructed 3D mesh – via a scale-shift transformation that minimises the least squares error to the mesh's depth rendered from the estimated camera viewpoint. We refer to this as the *aligned depth map*.

Given an observation $o_n = (f_n, m_n)$, we then assign a depth d_n to observation o_n corresponding to the centroid of the 2D mask m_n on the aligned depth map. We take the object's 2D location in frame f_n, depth relative to the camera d_n, and camera pose C_{f_n}, and project the observation to the fixed 3D world coordinate, such that:

$$[X_n, Y_n, Z_n]^T = \mathcal{C}_{f_n} \begin{bmatrix} d_n K^{-1} [x_n, y_n, 1]^T \\ 1 \end{bmatrix} \quad (1)$$

where K represents the camera's intrinsic parameters. We denote this 3D location as $l_n \in \mathbb{R}^3$. We visualise *lifting* to 3D in Figure 2. Note that we represent each observation as a point in 3D following previous works [16, 26]. These 3D observations are still partial and only on individual frames.

Visual features. In addition to the 3D location, we also compute visual features for each observation o_n which we need to match observations over time into 3D tracks. We denote this as $v_n = \Psi(f_n, m_n)$, where Ψ is a function that represents the visual feature extractor applied to the mask m_n on the frame f_n.

Lifted Visual Observations. We incorporate the 3D locations and visual features to give our set of partial observations $\mathcal{W} = \{w_n : n = 1, ..., N\}$ in the world coordinate frame, where $w_n = (f_n, l_n, v_n)$. We next describe how we match these observations over time to form 3D tracks.

3.2. Match and Keep Lifted Observations

Given the set of lifted observations, in this section we describe how to assign observations to consistent identities (*i.e.* track objects) across time. Object permanence dictates that objects do not actually *disappear* when occluded or are out of the egocentric camera's view – humans use spatial cognition to maintain the knowledge of where objects are.

We process the egocentric video E online, mimicking human spatial cognition: an object's location is tracked only after it is first encountered and this is when it is kept in mind.

Track definition. Each track $\mathcal{T}^j$ represents the set of observations belonging to the same object. We refer to the set of all tracks at time t as $\mathcal{T}_t$. A track has one 3D location at each time t, whether the object is in-sight or not, and we refer to this location by $L(\mathcal{T}_t^j)$.

Additionally, the track has an evolving appearance representation over time. It is calculated at time t using the visual appearance of the most recent γ visual features assigned to the track. Averaging visual features enhances representation robustness. Limiting the average to γ recent frames accounts for appearance changes over time (*e.g.* a bowl may be full, dirty, then clean). The track's appearance at time t is denoted as $V(\mathcal{T}_t^j)$.

Track initialisation. If an observation w_n represents a new, previously unseen object, i.e., is not matched to another track using the online matching described next, we initialise a new object track with this observation. We define an initialisation function $\mathcal{I}$, which initialises a new $\mathcal{T}^{J+1}$, where J tracks already exist, to the current 3D location and appearance of the observation w_n. As this is the first observation of the object, the track is projected back in time from the start of the video. $\forall t \leq f_n$:

$$\mathcal{I}(w_n) \rightarrow \mathcal{T}^{J+1} : L(\mathcal{T}_t^{J+1}) = l_n \text{ and } V(\mathcal{T}_t^{J+1}) = v_n \quad (2)$$

This reflects the common sense that objects do not magically appear out of thin air, so the first encounter of an object is an indication of its presence in that location earlier.

Track update. Once a track is initialized, its appearance and location are updated using new observations when available. The track update function $\mathcal{U}$ takes the track, observation and time as input:

$$\mathcal{U}(\mathcal{T}^j, w_n, t) \rightarrow L(\mathcal{T}_t^j) = l_n \text{ and } V(\mathcal{T}_t^j) = \mu(v_n, \mathcal{T}^j) \quad (3)$$

where μ calculates the mean of the past γ observations assigned to the track $\mathcal{T}^j$. If the track $\mathcal{T}^j$ is not assigned a new observation at time t then its representation remains unchanged: $\mathcal{U}(\mathcal{T}^j, \varnothing, t) \rightarrow \mathcal{T}^j_t = \mathcal{T}^j_{(t-1)}$.

Online Matching. We describe the process of forming tracks from online observations. We find the set of new observations at each t; $\mathcal{W}_t = \{w_n \quad \forall n : f_n = t\}$. Note that $\mathcal{W}_t$ is empty if there are no observations at time t.

Given the first frame with at least one observation, we initialise one track for each of these $\mathcal{T}_t = \{\mathcal{I}(w_n) \quad \forall w_n \in \mathcal{W}_t\}$. We next iterate over time and compare $\mathcal{W}_t$ to the set of trajectories at time $t-1$. Matching is based on a cost function using a combination of 3D distance and visual similarity, as in [36]. We model 3D similarity σ_L between an observation w_n and a track $\mathcal{T}^j$ by an exponential distribution, and visual similarity σ_V by a Cauchy distribution:

$$\sigma_L(w_n, \mathcal{T}^j) = \frac{1}{\beta_L} \exp\left[-D(L(\mathcal{T}^j_{t-1}), l_n)\right] \tag{4}$$

$$\sigma_V(w_n, \mathcal{T}^j) = \frac{1}{1 + \beta_V D(V(\mathcal{T}^j_{t-1}), v_n)^2} \tag{5}$$

where D is the Euclidean distance and β_L and β_V are relative weights for location and visual similarities.

We define the cost Φ of assigning an observation with an existing track as a combination of 3D and visual distance:

$$\Phi(w_n, \mathcal{T}^j) = -\log\left(\sigma_L(w_n, \mathcal{T}^j)\right) - \log\left(\sigma_V(w_n, \mathcal{T}^j)\right) \tag{6}$$

We then use a simple Hungarian algorithm as a robust method for associations as in [44, 59, 60]. Our matching algorithm ξ computes Φ between every observation in $\mathcal{W}_t$ and the tracks $\mathcal{T}_{(t-1)}$[1]. It returns a set of track assignments A_t for time t, where $A^j_t = w_n$ indicates that the track $\mathcal{T}^j$ is to be assigned the observation $w_n \in \mathcal{W}_t$:

$$A_t = \xi(\Phi, \mathcal{W}_t, \mathcal{T}_{t-1}) \tag{7}$$

We update tracks and initialise new tracks, such that:

$$\mathcal{T}^t \leftarrow \begin{cases} \mathcal{U}(\mathcal{T}^j, A^j_t, t) & \forall j \\ \mathcal{I}(w_n) & \forall w_n \in \mathcal{W}_t : \left(\nexists j : A^j_t = w_n\right) \end{cases} \tag{8}$$

By following the proposed online matching, we have an estimate of the 3D location for every object for which there is at least one observation.

3.3. LMK for object visibility and positioning

As a result of the spatial cognition enabled by the Lift-Match-and-Keep process, we are able to provide further information about the visibility of each object in relation to the camera wearer at time t. An object j can be *one* of:

- **In-sight**: if the corresponding track is assigned an observation at time t, *i.e.* $A^j_t \neq \varnothing$
- **Occluded**: if $L(\mathcal{T}^j_t)$ is within the field of view of the estimated camera C_t, but there is no corresponding observation ($A^j_t = \varnothing$). Note that without additional knowledge we cannot distinguish between missing observations and occlusion.
- **Out-of-view**: if $L(\mathcal{T}^j_t)$ is outside the field of view of the camera C_t.

An object may also be referred to as **Out-of-sight** if it is either out-of-view or occluded (*i.e.* in the camera's viewing direction but cannot be detected as it is behind or inside another object).

LMK also discloses the relative distance between the object and the camera-wearer or the static environment:

- **In-reach**: if the distance from object j to the camera's position at time t is within the camera wearer's near space η: $D(L(\mathcal{T}^j_t), C_t) \leq \eta$
- **Out-of-reach**: as in-reach, but if $D(L(\mathcal{T}^j_t), C_t) > \eta$.
- **Moved**: object j has moved *relative to the environment* between times t_1 and t_2 if $D(L(\mathcal{T}^j_{t_2}), L(\mathcal{T}^j_{t_1})) \geq \epsilon$, where ϵ is a minimum threshold (to account for small errors in camera and object positions).
- **Stationary**: as moved, but $< \epsilon$.

Note that an object may be both occluded but in-reach.

4. Experiments

Section 4.1, introduces our benchmark for the OSNOM task. Section 4.2 details baseline methods. Section 4.3 contains the main results and qualitative examples. Section 4.4 ablates LMK, including its capabilities for spatial cognition.

4.1. Benchmarking OSNOM

Dataset. We evaluate on long videos from the EPIC-KITCHENS [7] dataset. This dataset offers unscripted recordings of single participants – all object motions are thus a result of the camera wearer moving objects around. We select on 110 videos from EPIC-Kitchens, 12 minutes long on average, from 45 different kitchens, for a total of 25 hours of videos. We randomly select 10 validation videos for hyperparameter tuning and 100 videos for evaluation. These videos contain a total of 7.9M masks, which correspond to 2939 objects. We use the object semantic label only for calculating the ground truth for evaluation.

For most of our results, we use masks provided by VISOR [8], which are interpolations of ground-truth masks. This allows us to assess LMK's performance without accumulating errors from a object detector. For completion, we also ablate these results with the usage of a semantic-free detector [43] in Section 4.4.

Benchmark task. Due to the length of our videos, an exhaustive evaluation for every object from every frame is intractable. We thus select challenging key-frames $\mathcal{F}$ –

[1] A threshold for assignment cost is set to α

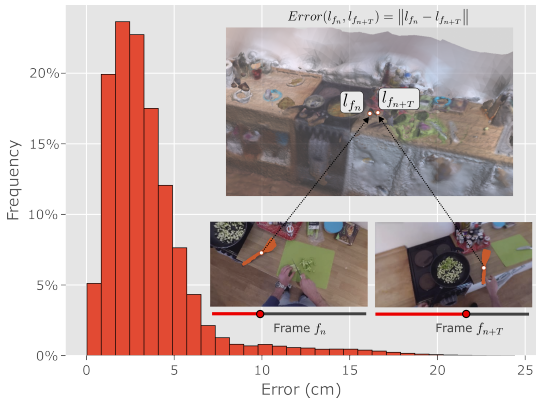

Figure 3. **3D Projection error.** Distribution of Euclidean distance errors for the same object, at one location, comparing l_n to l_{n+T}.

these are frames with 3 or more objects being interacted with. Each frame $f \in \mathcal{F}$ includes objects that are in-sight and we wish to evaluate the methods' ability to correctly locate the 3D locations of these same objects over frames $f \pm \delta$. We compare the performance of different methods as δ increases. In total, we evaluate starting from $\mathcal{F} = 3299$ frames, locations at $603k$ frames and 2007 objects, averaging $49k$ frames and 20 objects per video. Our benchmark is publicly available for comparisons (see Project Webpage).

Ground truth locations. We use our 2D to 3D lifting approach presented in Section 3.1 as ground-truth locations. We quantitatively assess the error in these locations as follows. We select a random set of objects and the corresponding time segments when these are in the same location throughout the environment. The error between the projections from multiple views, for the same object in the same location, allows assessing our 3D locations. Our analysis (details in Appendix B) shows that the mean 3D error is 3.5cm, with 88% of all errors smaller than 6cm and 96% of all errors smaller than 10cm (Figure 3). Given these results, we find our lifting to be sufficiently accurate to be used as ground-truth locations. This also informs our metric, where we ensure our threshold for accepting assignments is sufficiently larger than the error noted here.

Evaluation metric. Traditional tracking metrics do not evaluate tracks when out of sight [2, 25, 39]. Thus, we define a metric called Percentage of Correct Locations (PCL), drawing inspiration from the Percentage of Correct Keypoints (PCK) [54] used to evaluate pose estimation, to evaluate the spatial alignment of objects. PCL considers a correct prediction at time t if the object is correctly identified at time t and its predicted 3D location is within a threshold R from the ground truth 3D location. As PCL is calculated throughout time, any lost tracks are captured in the metric.

For our main experiments, we use $R = 30\text{cm}$[2]. This reflects that a function of spatial cognition is to know the location of an object with sufficient precision in order to navigate to or obtain it [11, 50]. R is visualised and ablated.

[2]Half the standard width of a cupboard/cabinet which is 60cm/24inch

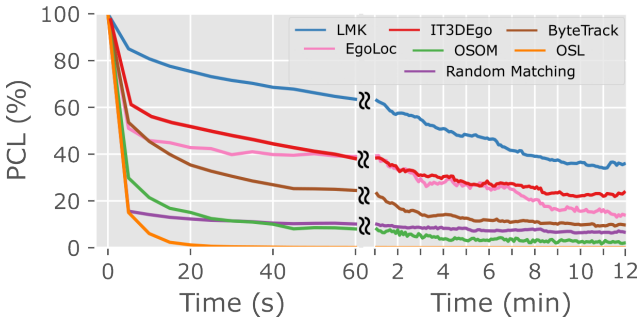

Figure 4. **OSNOM results.** PCL of LMK compared to baselines.

4.2. Experimental setup

Baselines. As no prior works have attempted the OSNOM task, we compare LMK against three naive baselines and three previous works adapted to the OSNOM task:

- **Random Matching**: each observation is randomly assigned either to an existing track or a new track, demonstrating the complexity of the data.
- Out of Sight, Lost **(OSL)**: objects are forgotten when they go out-of-view, so PCL is reported as 0 and their tracks are terminated. This baseline highlights the challenge in egocentric video, where objects move very frequently out of view soon after being first observed.
- Out of sight, out of mind **(OSOM)**: observations can only be assigned to tracks which are in-view. When a track goes out-of-view, PCL is reported as 0 and tracks are frozen until it is back in-view. This is an upper bound for tracking in the camera coordinate frame.
- **ByteTrack** [60]: a strong, recent 2D multi-object tracking method, widely used as a baseline [23, 33, 57]. Objects are tracked in 2D and then lifted in 3D using our lifting approach for evaluation.
- **EgoLoc** [26]: we adapt this SOTA VQ3D approach to OSNOM, to handle multiple objects. We use the same masks and lifting for fair comparison. EgoLoc's weighted averaging over all past observations fails for OSNOM because objects change position, so instead we take the most recent match.
- **IT3DEgo** [61]: As this paper uses ground truth depth which is not available in our RGB sequences, we instead run the 2D tracking using their public code, then lift the tracked objects to 3D using our approach (Sec 3.1).

Implementation details. For appearance features Ψ, we use a DINO-v2 [31]. We crop each mask, scale to 224×224 and pass to the backbone. We set $\alpha = 10$, $\gamma = 100$, $\beta_L = 13$ and $\beta_V = 2$ (chosen on the validation set). We compute meshes in advance, which takes 5 hours on average on one 2080Ti per video. Then for online tracking, DINOv2 operates at 30 FPS and lifting-to-3D at 200FPS on one P100. LMK runs at 1000fps on a single CPU core.

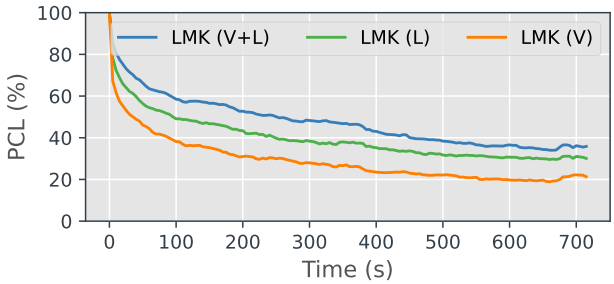

Figure 5. **Effect of visual appearance and location**. PCL for visual features (V), location features (L), or both (V+L).

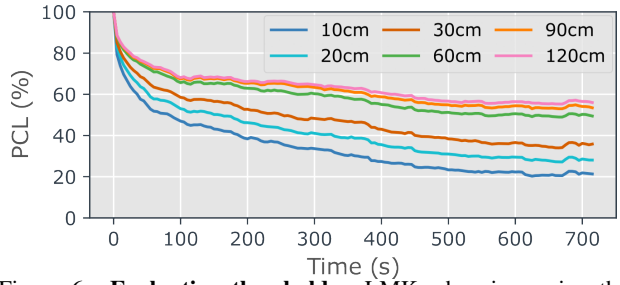

Figure 6. **Evaluation thresholds.** LMK when increasing the PCL threshold R - the maximum distance between predicted and ground truth 3D locations considered successful.

4.3. Results

Results on the OSNOM task for LMK, compared to the baselines, are shown in Figure 4. The average PCL (y-axis) over the whole dataset is reported for each 5s evaluation interval (shown on the x-axis), with standard deviation shaded. We show performance over both short (0-60 seconds) and long (1-12 minutes) timescales. Over time, the complexity of matching observations increases as more objects are being interacted with and tracked. This is reflected in performance decreasing for all methods over time.

LMK presents a significant improvement over all baselines. The rapid drop in performance in OSOM and OSL shows the challenge of egocentric footage, where the constantly moving person causes objects to go out of view frequently. When objects are tracked as long as in-view (OSL baseline), performance goes to zero just after 20s, showing that objects are quickly lost from sight. The OSOM baseline shows that only considering objects in-view, without 3D world coordinates and object permanence, is insufficient for the OSNOM task (is worse than random). LMK significantly outperforms ByteTrack, EgoLoc, and IT3DEgo. ByteTrack and IT3DEgo rely on 2D frames, while EgoLoc loses tracking quickly by comparing to initial appearances. In contrast, LMK tracks across consecutive frames, handling appearance changes from orientation or occlusion and leveraging 3D locations for robust matching.

4.4. LMK Ablation

Effect of visual appearance and location. LMK assigns observations to tracks based on appearance and location similarity. Figure 5 shows the effect of only visual appearance (V) and only location (L) compared to the default of both (V+L). Their combination shows improvements (mean +19% over V, +8% over L), highlighting that appearance and location are complementary. Appearance is good for frame-to-frame assignment, and location is particularly helpful for objects in motion, occluded and for reassigning objects when they reappear.

Accuracy at different radii. All our experiments set the PCL threshold, $R = 30\text{cm}$. Figure 6 also shows results when this is set to $R = 10/20/60/90/120\text{cm}$. As expected, PCL increases as R increases.

Detections. We used annotations from VISOR [8] as 2D masks. This avoids compounding detector error when evaluating the error of 3D location estimation, which is our primary task. In Figure 7 we show an ablation using detections from [43]. This model provides semantic-free bounding boxes of active objects, which we use as input to LMK and the best performing baseline EgoLoc. LMK still outperforms EgoLoc by a large margin.

LMK for spatial cognition. Figure 8 shows performance of LMK on the object states defined in Section 3.3. For each combination of (In-reach[3], Out-of-reach), (In-sight, Occluded, Out-of-view), we report the total number of ground truth objects and the number LMK correctly locates over a 1 minute interval. After 1 minute of objects being interacted with, LMK is still able to determine their locations, with an average accuracy of 72%. Additionally, LMK obtains 82% on objects which are out-of-reach and out-of-view.

We also investigate the ability of LMK to track objects going out- then back in-view (*i.e.* reappearing) within 10 minutes (Figure 9). LMK, matching using 3D locations, shows considerable performance improvement.

Qualitative results. Figure 10 shows the predicted locations of a couple of objects at discrete time scales. In Figure 11, we show 3D trajectories of objects as they are moved around by the camera wearer. For example, we show the trajectory of the *salt bottle* from being in the hand (pouring salt), placed on the countertop and eventually returned to a lower cupboard, while the *cup* ends on a hanger. In all cases, LMK is capable of accurately tracking objects when when static (on surfaces) and when moving (in-hand).

We include examples of failure cases in Appendix D.

5. Conclusion

In this paper, we introduced the task of "Out of Sight, Not Out of Mind" (OSNOM) for egocentric video with partial object observations. It evaluates 3D tracking performance

[3] We use a reachable threshold $\eta = 70\text{cm}$

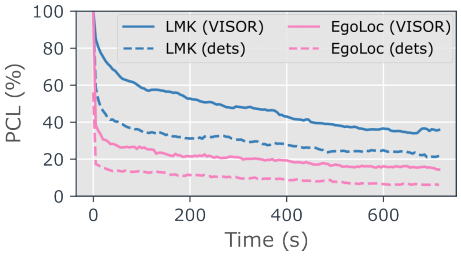

Figure 7. **Detections.** LMK and EgoLoc baseline [26] on both visual and location features (V+L) when using VISOR annotations *vs* using object detections from [43].

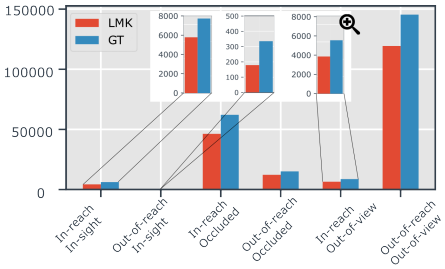

Figure 8. **LMK for spatial cognition.** Number of objects correctly located by LMK for each (In-reach, Out-of-reach) and (In-sight, Occluded, Out-of-view) combination.

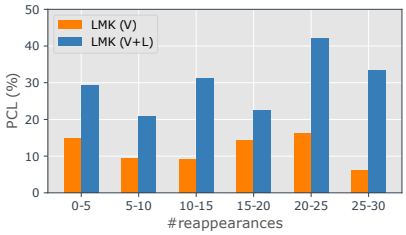

Figure 9. **Effect of reappearing.** Evaluation is performed over 10 minutes, for LMK with visual appearance (V) and the combination of visual appearance and location (V+L).

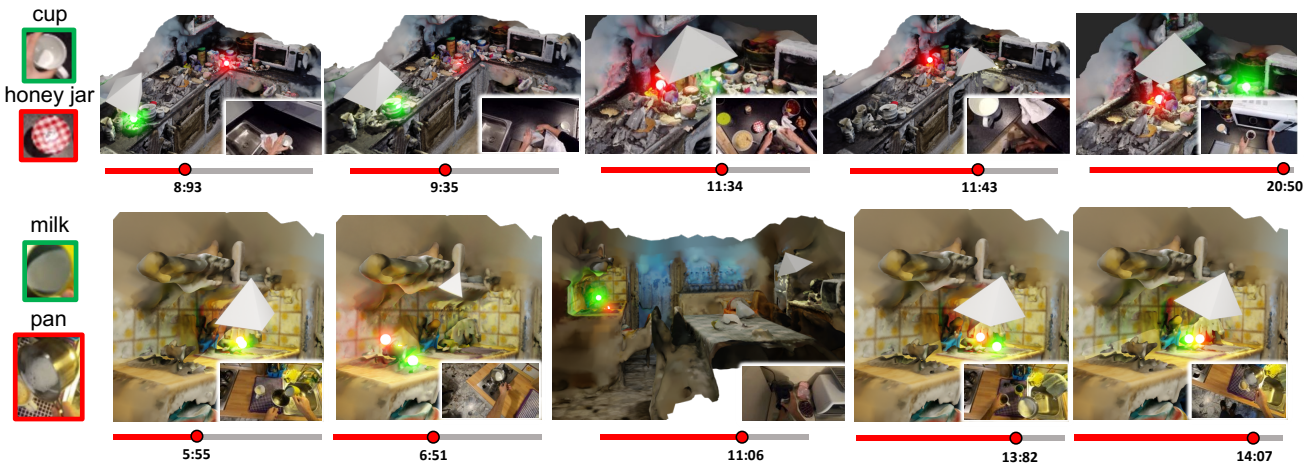

Figure 10. **3D location prediction.** Predicted 3D locations (Neon dots) of two objects (left) over multiple times with frame insets (right). Note how object locations are accurately kept in mind, even when the camera-wearer is far away (bottom middle).

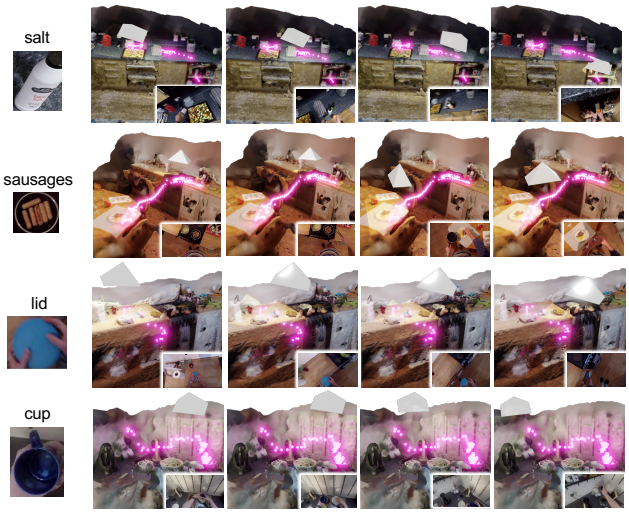

Figure 11. **Trajectory prediction** for objects in motion. Neon dots show correctly predicted 3D positions with corresponding camera views. Objects are accurately located both when static (on surfaces) and when moving (in-hand).

of active objects when they are both in- and out-of-sight.

We introduced a very strong baseline: Lift, Match and Keep (LMK), a method which *lifts* partial 2D observations in camera coordinates to 3D world coordinates, *matches* them over time using visual appearance and 3D location, and *keeps* them in mind when they go out of sight. Results on long videos from EPIC-Kitchens show LMK delivers good results over both short (64% after tracking for 1 minute) and long (37% after 10 minutes) timeframes, and that maintaining 3D world location is critical when objects go out-of-view. LMK outperforms recent works, strong 2D tracks and naive baselines by a big margin. For future work, we will investigate whether LMK can help track objects that undergo state changes, and explore shared 3D object tracks between multiple ego- and exo-centric cameras.

Acknowledgments. Research at Bristol is supported by EPSRC Fellowship UMPIRE (EP/T004991/1) & PG Visual AI (EP/T028572/1).We particularly thank Jitendra Malik for early discussions and insights on this work. We also thank members of the BAIR community and the MaVi group at Bristol for helpful discussions. This project acknowledges the use of University of Bristol's Blue Crystal 4 (BC4) HPC facilities.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *International Conference on Computer Vision*, 2019. 3

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[3] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *International Conference on Computer Vision*, 2009. 3

[4] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006. 2

[5] Siu-Wing Cheng, Tamal Krishna Dey, Jonathan Shewchuk, and Sartaj Sahni. *Delaunay mesh generation*. CRC Press Boca Raton, 2013. 3

[6] Giorgia Committeri, Gaspare Galati, Anne-Lise Paradis, Luigi Pizzamiglio, Alain Berthoz, and Denis LeBihan. Reference frames for spatial cognition: different brain areas are involved in viewer-, object-, and landmark-centered judgments about object location. *Journal of cognitive neuroscience*, 16(9):1517–1535, 2004. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022. 2, 3, 5

[8] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-Kitchens VISOR benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 5, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[11] Roger M Downs and David Stea. *Image and environment: Cognitive mapping and spatial behavior*. Transaction Publishers, 1973. 2, 6

[12] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 3

[13] Shane Gilroy, Edward Jones, and Martin Glavin. Overcoming occlusion in the automotive environment—a review. *IEEE Transactions on Intelligent Transportation Systems*, 22 (1):23–35, 2019. 3

[14] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *International Conference on Computer Vision*, 2021. 2

[15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision*, 2023. 3

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Computer Vision and Pattern Recognition*, 2022. 2, 3, 4

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[18] Shengyu Hao, Wenhao Chai, Zhonghan Zhao, Meiqi Sun, Wendi Hu, Jieyang Zhou, Yixian Zhao, Qi Li, Yizhou Wang, Xi Li, and Gaoang Wang. Ego3DT: Tracking every 3d object in ego-centric videos. In *ACM Multimedia 2024*, 2024. 2, 3

[19] Yan Huang and Irfan Essa. Tracking multiple objects through occlusions. In *Computer Vision and Pattern Recognition*, 2005. 3

[20] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *International Conference on Computer Vision*, 2023. 3

[21] Tarasha Khurana, Achal Dave, and Deva Ramanan. Detecting invisible people. In *International Conference on Computer Vision*, 2021. 3

[22] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. *International Conference on 3D Vision*, 2023. 3

[23] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18963–18973, 2024. 6

[24] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*. Springer, 2022. 2

[25] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129: 548–578, 2021. 6

[26] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Localizing objects in 3d from egocentric videos with visual queries. *International Conference on Computer Vision*, 2023. 2, 4, 6, 8

[27] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Computer Vision and Pattern Recognition*, 2022. 3

[28] M Keith Moore and Andrew N Meltzoff. Object permanence after a 24-hr delay and leaving the locale of disappearance: the role of memory, space, and identity. *Developmental Psychology*, 40(4):606, 2004. 2

[29] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Computer Vision and Pattern Recognition*, 2020. 2

[30] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egocentric scene context for human-centric environment understanding from video. *Advances in Neural Information Processing Systems*, 2023. 2

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 6, 1

[32] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *International Conference on Computer Vision*, 2023. 3

[33] Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19004, 2024. 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 1

[35] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *Advances in Neural Information Processing Systems*, 2021. 3

[36] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Computer Vision and Pattern Recognition*, 2022. 3, 5

[37] Siddharth Ravi, Pau Climent-Perez, Théo Morales, Carlo Huesca-Spairani, Kooshan Hashemifard, and Francisco Flórez-Revuelta. Odin: An omnidirectional indoor dataset capturing activities of daily living from multiple synchronized modalities. In *Computer Vision and Pattern Recognition*, 2023. 3

[38] Xuanchi Ren, Tao Yang, Li Erran Li, Alexandre Alahi, and Qifeng Chen. Safety-aware motion prediction with unseen vehicles for autonomous driving. In *International Conference on Computer Vision*, 2021. 3

[39] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*. Springer, 2016. 6

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition*, 2016. 3

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 3

[42] Aviv Shamsian, Ofri Kleinfeld, Amir Globerson, and Gal Chechik. Learning object permanence from video. In *European Conference on Computer Vision*, 2020. 3

[43] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 5, 7, 8

[44] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 5

[45] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Computer Vision and Pattern Recognition*, 2023. 3

[46] Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[47] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *International Conference on Computer Vision*, 2021. 3

[48] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea

Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems*, 2023. 3

[49] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *Computer Vision and Pattern Recognition*, 2023. 3

[50] David Ed Waller and Lynn Ed Nadel. *Handbook of spatial cognition.* American Psychological Association, 2013. 2, 6

[51] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[52] Kelvin Wong, Yanlei Gu, and Shunsuke Kamijo. Mapping for autonomous driving: Opportunities and challenges. *IEEE Intelligent Transportation Systems Magazine*, 13(1):91–106, 2020. 3

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *Computer Vision and Pattern Recognition*, 2024. 4

[54] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 6

[55] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Computer Vision and Pattern Recognition*, 2023. 3

[56] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Computer Vision and Pattern Recognition*, 2022. 3

[57] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, 2022. 6

[58] Jeroen Zewald and Ivo Jacobs. Object permanence. In *Encyclopedia of animal cognition and behavior*, pages 4711–4727. Springer, 2022. 2

[59] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021. 5

[60] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*. Springer, 2022. 2, 3, 5, 6

[61] Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3d scenes from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21933–21944, 2024. 2, 3, 6