Egocentric Prediction of Action Target in 3D

Yiming Li[1,*], Ziang Cao[2,*], Andrew Liang[1], Benjamin Liang[1], Luoyao Chen[1], Hang Zhao[3], Chen Feng[1,†]
[1]New York University [2]Tongji University [3]Tsinghua University

https://ai4ce.github.io/EgoPAT3D/

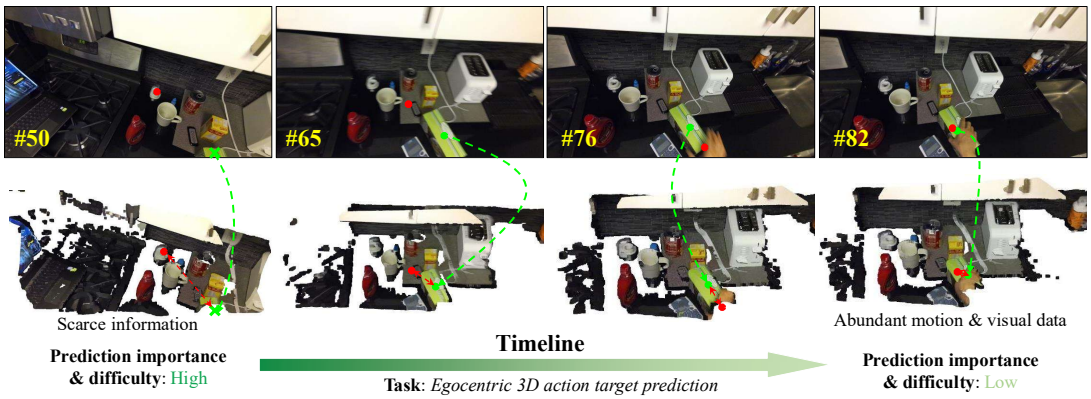

Figure 1. **Illustration of the proposed egocentric prediction task**. The predicted and ground truth target locations of an action (grasping) sequence are visualized by red and green dots. In the early stage, only scarce information is available, thus it is more challenging to achieve reliable prediction. Over time, more visual and motion cues are accumulated, making it easier to predict the action target. Note that the prediction is updated in each frame and the 3D action target in each frame is expressed in the coordinate system of that frame.

Abstract

We are interested in anticipating as early as possible the target location of a person's object manipulation action in a 3D workspace from egocentric vision. It is important in fields like human-robot collaboration, but has not yet received enough attention from vision and learning communities. To stimulate more research on this challenging egocentric vision task, we propose a large multimodality dataset of more than 1 million frames of RGB-D and IMU streams, and provide evaluation metrics based on our high-quality 2D and 3D labels from semi-automatic annotation. Meanwhile, we design baseline methods using recurrent neural networks and conduct various ablation studies to validate their effectiveness. Our results demonstrate that this new task is worthy of further study by researchers in robotics, vision, and learning communities.

1. Introduction

Egocentric vision, which parses images from a wearable camera capturing a person's visual field, has been an important area in robotics and computer vision due to its wide applications, *e.g.*, virtual reality [23], human-robot interaction [10], and social robotics [39]. In recent years, a variety of egocentric datasets and benchmarks have been established [8, 11, 19, 46, 52], and researchers have proposed various methods around the egocentric scene understanding, such as action recognition which summarizes an egocentric video clip into a certain action category [25, 51], and action anticipation which infers future action types (without location information) based on the historical information [16, 36]. However, a fundamental egocentric vision problem remains underexplored: *how to anticipate the future target location of someone's object manipulation action in 3D space*? This is crucial for more safe and effective planning and control of robots to collaborate with human.

Basically, cognitive robots that need to interact with humans should be able to expect target locations of human actions at an early stage, allowing robots to compute appropriate reactions. For example, when a person with upper-limb neuromuscular diseases is trying to grasp an object, a wearable exoskeleton should comprehend the human's intended target location before the grasping is completed, so the robot can plan its motion smoothly to help achieve the goal. Additionally, any computational latency in the prediction algorithm implementation could also be compensated

*Equal contribution
†The corresponding author is Chen Feng. cfeng@nyu.edu

if the robot is able to predict the action target ahead of time.

Thus, we establish a new dataset, EgoPAT3D, for egocentric prediction of action target in 3D. Our motivation lies in three aspects: **(1)** understanding the target locations of human actions is of significance to human-robot interaction [27], **(2)** prediction in 3D instead of 2D space facilitates the robot planning and control, and **(3)** there are no public datasets for egocentric action target prediction in 3D. To this end, we initiate the first study of *egocentric 3D action target prediction*, which processes an egocentric sensor stream as an online signal and formulates action understanding as a continuous update for the target location. In summary, the main purpose of this study is to anticipate the 3D target locations of human actions as early as possible, to *compensate for any latency* and *support timely reactions* of robots.

Technically, it might seem too difficult if not impossible to ask the machine to accurately anticipate human intention locations especially at an early stage when there is very limited information. However human behaviors while attaining goal locations have certain distinct properties such as *eyes are faster than hands* [27]: when we try to grasp an object, we firstly search for the object in our visual field before reaching out for it. This phenomenon indicates that human intention locations could be anticipated based on the information of both *visual perception* and *head motion*. Therefore, our dataset is multimodality, including RGB and depth images, and inertial measurement unit (IMU) data, which are all recorded by a helmet-mounted Azure Kinect RGB-D camera. In each recording, the camera wearer reaches for, grabs, and moves objects randomly placed in a household scene. Each recording features a different configuration of household objects within the scene. To annotate action targets less laboriously, we employ an off-the-shelf hand pose estimation model to localize the hand center which is used to denote the target location.

In order to solve this novel task of *egocentric 3D action target prediction*, we propose a simple baseline approach on top of recurrent neural networks (RNNs) in conjunction with both visual and motion features. To summarize, our main contributions are as follows:

- We initiate the first study of the egocentric action target prediction in 3D space.
- We build a novel EgoPAT3D dataset and propose new evaluation metrics for this new egocentric vision task.
- We design a simple baseline method to achieve continuous prediction of action target, and comprehensively benchmark the performance.
- We open source all the code and dataset for reproducibility and future improvements.

2. Related Work

Egocentric datasets and benchmarks. Egocentric videos provide a wealth of knowledge on how humans see and interact with their surroundings, which is vital to understanding human behavior. The research in egocentric vision has been rapidly advanced owing to the development of wearable devices as well as egocentric datasets. In recent years, the scale of datasets has been gradually increased, and both scenes and annotations in egocentric scenarios have been enriched. For example, 2D object bounding boxes are provided to facilitate a variety of 2D computer vision tasks [7, 9, 11, 46, 50]. Gaze measurements are supplemented to help understand the human intention in the image space [14, 24, 33, 60]. Hand annotations are provided as a useful information to understand human-object interaction [2, 19]. Thanks to these well-built datasets, various egocentric vision tasks have been proposed and studied such as action recognition [25, 34, 51], action anticipation [16, 20, 36], video summarization [30, 38, 55], hand-object interaction parsing [3, 6, 41], social interaction analysis [13, 42, 56], and egocentric object detection and tracking [2, 11, 32]. Despite numerous efforts to promote the development of egocentric vision, *most datasets and tasks focus solely on 2D computer vision without 3D data*. Ego4D [21], a large-scale egocentric video dataset partially containing audio, mesh, stereo, and eye gaze information, was recently presented, and it also proposed five benchmarks centered on episodic memory, hands and objects, audio-visual diarization, social interactions, and activity forecasting. *Despite its unprecedented scale and diversity, online 3D target location prediction in egocentric views remains insufficiently investigated.*

Egocentric future prediction. In literature, there are substantial works in egocentric action recognition [1], video summarization [40], hand analysis [3], and future prediction [49]. Here we only review the most relevant work, egocentric future prediction, which is a relatively new research area. Egocentric prediction of human activities, targets, and trajectories has wide applications such as assistive technologies [43], trajectory planning [44], multimedia [35], and robotics [27]. There are mainly three streams in egocentric future prediction: **(1)** action anticipation, **(2)** region prediction, and **(3)** trajectory forecasting. The first two problems are intensively studied yet there are scarce works regarding the third one. Action anticipation aims to generate an action label given a historical video clip, and various datasets such as EPIC-KITCHENS [7, 8] and EGTEA Gaze+ [33] which can support action anticipation have promoted the research in this topic [17, 18, 26, 59]. Region prediction is to predict a 2D region on the image which will cover the human intended location in the future, and the target regions are denoted by object bounding boxes [5, 12], human-object interaction hotspots [36], or future gaze locations [57, 58]. Trajectory prediction attempts to forecast the future foot trajectories of humans [4, 44]. For example, Park *et al.* proposed to generate plausible future trajecto-

Table 1. Comparison of datasets which can support research in **egocentric future prediction**. Prediction tasks are divided into three kinds: **(1)** action anticipation, **(2)** region prediction, and **(3)** trajectory/target prediction. The datasets listed in the **upper section** provide 2D cues, and the datasets in the **lower section** include 3D information such as depth and inertial measurement unit. N/P denotes Not Provided.

Dataset	Prediction Task	Scenarios	Device	Modality	Annotation	# of Frame	# of Env.	Year	Public
ADL [46]	Region	Manipulation	GoPro	RGB	Action/Object	1.0M	20	2012	✓
GTEA Gaze+ [14]	Region	Manipulation	SMI	RGB/Audio	Hand/Gaze	0.4M	1	2012	✓
Daily Intentions Dataset [53]	Action	Manipulation	Fisheye Len	RGB/IMU	Action	N/P	N/P	2017	✓
EPIC-KITCHENS-50 [7]	Action/Region	Manipulation	GoPro	RGB/Audio	Action/Object	11.5M	32	2018	✓
MAD [15]	Action	Manipulation	N/P	RGB/Force	Action	N/P	N/P	2018	✗
ATT [60]	Region	Manipulation	N/P	RGB	Gaze	217.0K	N/P	2018	✗
EPIC-Tent [24]	Region	Manipulation	GoPro/SMI	RGB/Audio	Gaze	1.2M	1	2019	✓
100DoH [50]	Region	Manipulation	YouTube	RGB	Hand	27.3K	N/P	2020	✓
EPIC-KITCHENS-100 [8]	Action/Region	Manipulation	GoPro	RGB/Audio	Action/Object	20.0M	45	2020	✓
EGTEA Gaze+ [33]	Region	Manipulation	SMI	RGB/Audio	Hand/Gaze	2.4M	1	2021	✓
MECCANO [47]	Region	Manipulation	SR300	RGB	Object	0.3M	N/P	2021	✓
Ego4D [21]	Action/Region	Manipulation	GoPro	RGB	Action/Object	N/P	N/P	2021	✓
KrishnaCam [28]	Trajectory	Walking	Cellphone	IMU/GPS	Trajectory	7.6M	N/P	2016	✓
EgoMotion [44]	Trajectory	Walking	GoPro Stereo	RGB-D	Trajectory	65.5k	26	2016	✓
Ego4D [21]	Trajectory	Walking	Stereo	RGB-D	Trajectory	N/P	N/P	2021	✓
EgoPAT3D (ours)	Target	Manipulation	Azure Kinect	RGB-D/IMU	Target	1M	15	2021	✓

ries of human ego-motion in egocentric stereo images [44]. Rhinehart *et al.* used online inverse reinforcement learning to forecast a person's walking destination and action in a 3D map [48]. The recently-proposed Ego4D [21] developed a unified benchmark to evaluate the progress in egocentric future prediction including all the three prediction tasks (action/region/trajectory). However, the online prediction of action target in the 3D space still remains to be studied.

Remark. Existing datasets related to egocentric future prediction are summarized in Table 1. In summary, most research in egocentric prediction have concentrated on the action category or 2D image region, and a few works have studied the egocentric 3D trajectory prediction in the walking scenarios. However, 3D action target prediction in the manipulation scenarios with rich hand-object interactions is still underexplored. Actually, target prediction is a special case of trajectory forecasting: the former only computes the end points of the trajectories. Yet in the manipulation scenarios, it is often infeasible to obtain complete hand trajectories because human hands often locate outside of the egocentric view. Therefore, we focus on predicting the 3D target location of an object manipulation action, which is desirable in robot planning and control.

3. Egocentric Action Target Prediction in 3D

In this section, we define the problem of egocentric 3D action target prediction, discuss the challenges of the task, introduce the evaluation metrics for the proposed task, and present our simple baseline method.

3.1. Problem Formulation

Many prior works in egocentric future prediction consider an offline mode, *i.e.*, they require a fixed-length historical sequence to predict the future action or target region. Differently, we consider a more realistic online mode, *i.e.*, our task requires online prediction based on variable-length historical information. Meanwhile, the difficulty level of our task varies in different temporal stages, *e.g.*, in the early stage, the task is very challenging because **(1)** there is scarce information, and **(2)** the hands which can serve as crucial cues are often outside of the view. In contrast, the temporal information becomes rich and the hands are often clearly visible in the late stage, thus the task becomes easier over time. An example sequence is shown in Fig. 1.

Notation. The colored point cloud at frame t denoted by $\mathbf{X}_t \in \mathbb{R}^{N_t \times 6}$ is represented as a set of 3D points $\{\mathbf{x}_t^n | n = 1, 2, ..., N_t\}$, where the n-th point of frame t written as $\mathbf{x}_t^n \in \mathbb{R}^6$ is a vector of its Euclidean coordinate as well as RGB value (x, y, z, r, g, b), and N_t denotes the number of points at frame t. The IMU data at frame t denoted by $\boldsymbol{\theta}_t \in \mathbb{R}^6$ is composed of the body-frame angular velocity $\boldsymbol{\omega}_t \in \mathbb{R}^3$ and linear acceleration $\boldsymbol{\alpha}_t \in \mathbb{R}^3$. The 3D target location at frame t represented in the corresponding coordinate is denoted by $\boldsymbol{y}_t \in \mathbb{R}^3$.

For a clip with length T, the prediction will be executed T times to achieve continuous update for the action target in this clip. Note that the action target $\{\boldsymbol{y}_t | t = 1, 2, ..., T\}$ within a clip is the same point in the world coordinate , yet has different values because they are represented in different local coordinates which depend on the head poses.

Definition. Given historical colored point cloud streams $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_t\}$ and IMU streams $\boldsymbol{\theta}_{1:t} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_t\}$, we aim to predict the 3D target location $\boldsymbol{y}_t$ at each frame. From the machine learning perspective, f denotes a model which takes visual and motion sensor streams as input, and can output an estimation of future target locations: $\boldsymbol{o}_t = f(\mathbf{X}_{1:t}, \boldsymbol{\theta}_{1:t})$. We seek an optimal f to predict $\boldsymbol{o}_t$ as close to $\boldsymbol{y}_t$ as possible. Meanwhile, we prefer to generate accurate predictions as early as possible.

Challenges. There are three major challenges in EgoPAT3D. **(1)** *Multi-modal information fusion*: the camera motion and visual information in a data stream have different importance for action target prediction at different ac-

tion stages. Intuitively, the head motion appears to be more beneficial in the early action stage, while the visual appearance seems more useful in the late stage when the hand position start to be observed and can be exploited. It is nontrivial to effectively fuse these two types of information to achieve reliable prediction. **(2)** *Early prediction*: achieving early stage prediction with high precision is usually more valuable for downstream applications such as robot control. But this is also difficult since information from the initial stage of an action is insufficient. **(3)** *3D workspace*: predicting in 3D space further increases the task difficulty compared to predicting on a 2D image.

Evaluation metrics. **(1)** *Temporal-aware evaluation:* we divide each temporal window (an action clip) into ten stages, and calculate average center location error (CLE) for the predictions in each stage, so that we can observe the tendency of the prediction precision over time. **(2)** *Early prediction evaluation:* we employ the prediction precision (CLE) when observing only the beginning 10%, 20%, 30%, 40%, and 50% of the action sequence, to assess the early prediction capability. **(3)** *Overall evaluation:* we linearly weight the prediction errors based on the temporal stages: the prediction errors at the early stages are strongly penalized, and we can compute an overall score using temporal-aware weighted sum of the errors at different stages.

3.2. Baseline Method

As mentioned in Section 3.1, egocentric 3D action target prediction is a very challenging task. We propose a simple baseline method which uses two backbone networks separately for multimodality representation learning, followed by utilizing concatenation to achieve multimodality feature fusion, and we employ a recurrent neural network (RNN) to achieve continuous update for the 3D action target. The main workflow is presented in Fig. 2.

Visual and motion features encoding. We use a visual feature extractor denoted by ψ which is based on a classic point cloud backbone PointConv [54], to encode the visual features $\boldsymbol{v}_t = \psi(\mathbf{X}_t)$. Besides, we use a motion feature extractor denoted by ϕ which is based on multilayer perceptron (MLP) to encode the motion cues $\boldsymbol{m}_t = \phi(\boldsymbol{\theta}_t)$. After the feature encoding, the features of two modalities are concatenated and fed into another MLP to obtain the fused features $\boldsymbol{u}_t = \text{MLP}(\text{Cat}(\boldsymbol{v}_t, \boldsymbol{m}_t))$.

Online 3D action target prediction. We divide the 3D space into grids and predict confidence value for each grid. We use a RNN to encode the sequential information and achieve online prediction, for example, at frame t, the score vector $\boldsymbol{s}_t \in \mathbb{R}^N$ for N-dimensional x-grids is computed by $\boldsymbol{s}_t = \text{RNN}(\boldsymbol{u}_t, \boldsymbol{h}_{t-1})$ (y and z directions are the same, so they are bypassed for simplicity), where $\boldsymbol{h}_{t-1}$ is the learned hidden representation. RNN is able to learn both long- and short-term dependency in historical sensory streams which is desirable in our task.

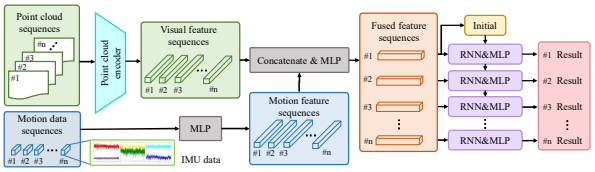

Figure 2. **Workflow of our baseline method.** The visual and motion features are separately extracted by two backbone networks, and then fused and fed into the RNN-based prediction module for the future action target localization.

Training objective and loss function. Since the adjoined grids are highly correlated compared to the general classification task, adopting the classic classification loss function (cross entropy loss) is not a promising choice, as demonstrated later in Section 5.2. Therefore, we redesign the loss function to utilize the dependencies between adjoined grids: we select the grid with a higher confidence score, and use the score to weight its importance. The comprehensive experiments prove that our loss, named as truncated weighted regression loss (TWRLoss), is more robust than the general one in this task which solely considers the single point with the highest confidence value. Mathematically, $\boldsymbol{g} \in \mathbb{R}^N$ denotes x-grids (we normalize the grid coordinate from -1 to 1), and the prediction score in the n-th x-grid is denoted by $\boldsymbol{s}_t[n] (n = 0, 1, ..., N-1)$. Let $\boldsymbol{m}_l \in \mathbb{R}^N$ represent the binary mask for x-grids at frame t to filter out grids with lower scores, and $\boldsymbol{m}_l[n]$ denote the binary value for the n-th x-grid, then the masked score in x-grids is calculated by:

$$\hat{\boldsymbol{s}}_t = \boldsymbol{m}_l \odot \boldsymbol{s}_t,$$
$$\boldsymbol{m}_l[n] = \begin{cases} 1, & n \in \{j \mid \boldsymbol{s}_t[j] > \gamma\} \\ 0, & n \in \{j \mid \boldsymbol{s}_t[j] \leq \gamma\}, \end{cases} \tag{1}$$

where $\odot$ denotes the element-wise dot product, and γ is a threshold to filter out the grids with low prediction scores. It is set to 0.5 in this method. Then the estimated target location's x-coordinate is calculated by: $p_t \in \mathbb{R} = \hat{\boldsymbol{s}}_t^T \boldsymbol{g}$. Since the prediction difficulty is different as time goes on, the prediction error at different stages should be penalized differently, that is, we assign different weights to losses at different stages: $\mathcal{L} = \sum_{t=1}^{T} w_t (y_t - p_t)^2$, where y_t is the x-coordinate of the ground truth target point, and w_t is a linear weight from 2 to 1, *i.e.*, $w_t = 2 - \frac{t}{T}$.

4. EgoPAT3D Dataset

4.1. Raw Data Acquisition

The raw EgoPAT3D data was recorded by 2 participants in 15 commonplace yet diverse household scenes. The recordings feature environments such as various kitchen, bedroom, and bathroom spaces. The participants continuously re-arranged the objects within the scene. Note that object rearrangement is a well-known task in the robotics community [22, 29, 31]. We purposefully chose scenes in

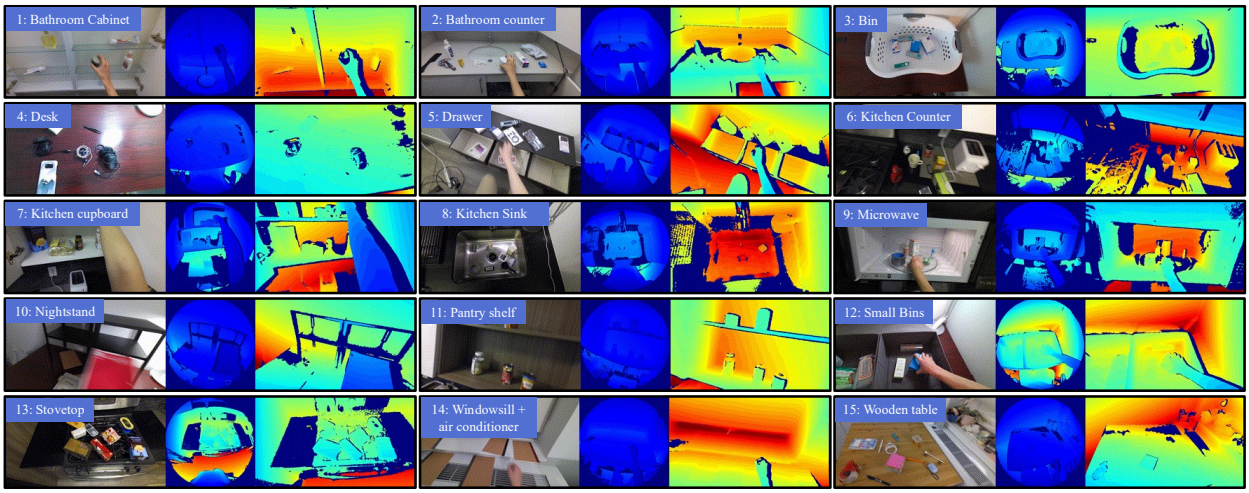

Figure 3. **Visualizations of scenes.** Left to right in three columns: RGB images, depth, depth transformed into RGB camera.

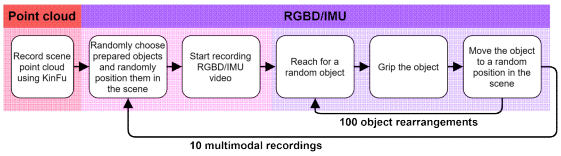

Figure 4. **Illustration of the data collection pipeline.** In each scene, we randomly choose and place objects in the scene and then conduct object rearrangement for one hundred times.

which hand-object manipulations, specifically grabbing and moving, often occur in everyday life (e.g. shelves, cabinets, counter tops, and other surfaces and fixtures where objects are stored or placed). All scenes are visualized in Fig. 3. The data collection procedure is demonstrated in Fig. 4.

The raw data is captured in 15 scenes, and consists of 10 RGB-D/IMU recordings per scene. Meanwhile, a point cloud of the default state of 15 scenes (objects removed from the environment) is included. Each recording features approximately 100 hand-object actions and 4 minutes of footage at 30 frames per second (FPS) for both color and depth streams. The total collection contains 150 recordings, 15 household scene point clouds, 15,000 hand-object actions, 600 minutes of raw RGB-D/IMU data, 0.9 million hand-object action frames, and 1 million RGB-D frames for the entire dataset. All RGB-D/IMU data was collected using Microsoft's Azure Kinect DK and Azure Kinect Recorder software. The recordings were created with the following settings: 3840x2160 (4K) color camera resolution, 512x512 depth camera resolution, 30 fps for both color and depth streams, and Wide field-of-view depth mode (WFOV) 2x2 binned depth recording mode to more closely simulate the large field-of-view (FOV) of human vision. Depth delay was set to 0, and IMU recording was turned on. Scene point clouds were generated using OpenCV KinectFusion (KinFu) for Azure Kinect.

Data was acquired in sessions, during which a participant would be asked to record all the raw data corresponding to a selected scene in our dataset. No monetary compensation was offered to the participants; all recordings were performed on a purely voluntary basis, and no personally identifiable information was present in the process. Participants wore a helmet with an attached front-facing Azure Kinect camera, angled downward to capture an egocentric field-of-view and reliably record the participant's head and hand movements in all recordings during the session. The scene would be cleared before each session by removing nearby visible objects that could be manipulated by hand. A large collection of random items that can be grasped by hand, including but not limited to those that might realistically appear in the scene in daily situations (i.e. cooking utensils, dried foods, and spices for kitchen scenes), were prepared as objects that could later be placed in random configurations in the scene prior to each recording during the session. At the start of the session, the participant would be directed to follow a set of instructions:

1. Record a point cloud of the scene using KinFu installed on a nearby laptop.
2. Adjust and wear helmet with mounted Azure Kinect so that the camera has a sufficient egocentric view of arm movements and hand-object interactions.
3. Arbitrarily pick and arrange several of the prepared items in the scene.
4. Stand or sit at designated location dependant on the scene, where any placed objects would easily be within field of view and arm's reach.
5. Start RGB-D/IMU recording using Azure Kinect Recorder installed on a nearby laptop.
6. Use a hand to re-arrange object in the scene 100 times.
7. Stop recording and remove objects from the scene.
8. Repeat steps 4-7 ten times for the scene, concluding the recording session.

4.2. Ground Truth Generation

Given a recording, we manually divide it into multiple action clips. To localize the 3D target in each clip, we use the following procedures. Firstly, we take the last frame of each clip based on the index provided by the manual division. Secondly, we use an off-the-shelf hand pose estimation model to localize the hand center in the last frame of each clip. Thirdly, we use colored point cloud registration to calculate the transformation matrices between the adjacent frames. Finally, for each clip, we transform the hand location in the last frame to historical frames according to the results of the third step, and the transformed locations can describe the 3D action target location in each frame's coordinate. Detailed procedures are presented as follows.

Manual clip division. We target short-term action target prediction, so we need to divide a long recording into multiple action clips such as reaching out for an object or placing an object. Specifically, we save the indexes of the first and the last frame for each clip, and such manual division is quite efficient: it takes around half an hour to manually annotate each recording. As shown in Fig. 5, most action clips have 10-40 frames. After obtaining the index of the last frame for each clip, we use an off-the-shelf hand pose estimation model to localize the hand center, which is considered as the ground truth target.

Hand pose estimation. For the last frame in each clip, 3D hand pose estimation is performed using Google's MediaPipe Hands python solution API, which first performs a single-shot palm detection task [37] before localization of 21 hand keypoints according to the MediaPipe hand landmark model. X and Y pixel coordinates of the keypoints were inferred this way, and the depth information (Z coordinate) of the hand is extracted from the corresponding depth frame transformed into color frame dimensions using the Azure Kinect SDK. Some hand pose estimation visualizations can be found in the supplementary.

Visual odometry. Since the camera (head) keeps moving when humans perform actions, the 3D target's coordinate is always changing although it is the same point in the world coordinate. We use colored Iterative Closest Point (ICP) [45] to compute the transformation matrices between two adjacent frames, which is usually called visual odometry. For each action clip, we can extract the hand location in the last frame to denote the target location for this action. Then we transform it into previous frames' coordinate system according to ICP. Therefore, the ground-truth action target in each frame could be generated. The distribution of the ground-truth target position is shown in Fig. 5.

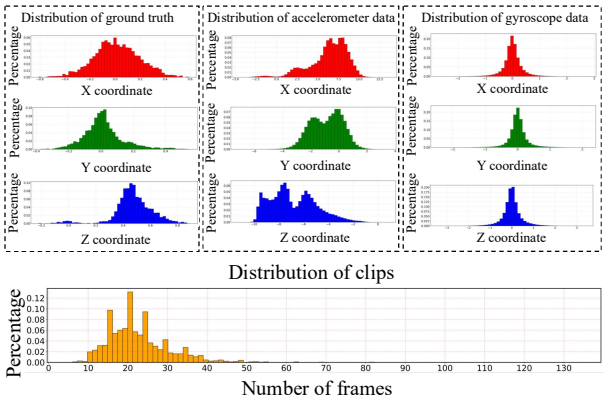

Figure 5. **Statistical properties of EgoPAT3D.** The distribution of 3D target locations and head motion (accelerometer and gyroscope) are visualized in xyz coordinate. The bottom figure shows the distribution of the number of frames in each action clip.

5. Experiments

5.1. Experimental Setup

Dataset preparation. We use 11 scenes in our experiments, *i.e.*, 5 seen scenes: bathroom cabinet, bathroom counter, drawer, kitchenCounter, nightstand, and 6 unseen scenes: bin, kitchen cupboard, microwave, stovetop, windowsill+air conditioner, and wooden table. Our training, validation, and test sets are composed of 1990, 358, 314 action clips respectively, and the unseen test set which is employed to test generalization ability contains 772 action clips from 10 scenes. The length of each action clip ranges from 6 frames to 133 frames (23 frames on average).

Implementation details. During the training of 30 epoches, the SGD optimizer is employed with an initial learning rate of 0.01 and a decay factor of 0.9 for every 5 epochs. Besides, the momentum, weight decay of SGD, and batch size are set to 0.9, 10^{-4}, and 8 respectively. In our baseline method, we adopt two RNN layers (LSTM or GRU) to exploit the temporal information. Our action target predictor is trained on NVIDIA GeForce RTX 3090 GPUs. We report the results on the test set.

5.2. Quantitative Results

The temporal-aware quantitative results are presented in Table 2. In the seen scenes, the error can be decreased from $\sim$ 25cm to $\sim$ 15cm, which is around the average length of an adult male's hand. Therefore, our baseline approach achieves a satisfactory performance in the task of egocentric prediction for action target. Meanwhile, there is still a remarkable gap to fill, so the proposed task is worthy of further investigation. The experimental results on unseen scenes validate the generalization ability of our baseline method, as shown in Table 3.

Discussions on features. To test the effect of different features on the performance, ablation studies on input fea-

Table 2. **Quantitative results of different predictors on seen scenes.** Note that the lower score represents better performance. VF denotes visual features, TF denotes transformation matrices, IMU includes angular velocities as well as linear accelerations, r denotes the reverse of linear loss weight function (from linearly decreasing to linearly increasing). Red, green, and blue fonts denote the top three performance.

Components	Overall (cm)↓	Early prediction (cm)					Late prediction (cm)				
		10%↓	20%↓	30%↓	40%↓	50%↓	60%↓	70%↓	80%↓	90%↓	100%↓
VF-r	19.21	23.67	22.02	20.58	19.27	18.16	17.29	16.69	16.50	16.42	16.63
VF	19.15	**23.22**	21.80	20.98	19.79	18.32	**17.29**	**16.59**	**16.08**	**16.03**	**16.24**
TF+IMU-r	**18.66**	**22.84**	**21.21**	**19.94**	**18.76**	**17.68**	**16.93**	**16.33**	**16.08**	**15.99**	16.24
TF+IMU	19.81	23.97	22.16	20.90	19.76	18.79	18.09	17.69	17.55	17.45	17.54
VF+IMU-r	20.87	24.70	23.40	22.37	21.08	19.83	19.12	18.67	18.39	18.31	18.41
VF+IMU	21.81	25.36	24.33	23.28	22.39	21.34	20.36	19.49	19.06	18.85	18.90
VF+TF-r	20.41	23.50	23.51	22.81	21.53	20.06	18.86	17.77	17.09	16.74	16.69
VF+TF	19.48	23.47	22.31	21.25	20.09	18.87	17.92	16.95	16.32	16.11	**16.18**
VF+TF+IMU-r	**18.97**	**23.02**	**21.11**	**20.01**	**19.03**	**18.13**	17.40	16.85	16.65	16.58	16.81
VF+TF+IMU	**18.61**	23.73	**21.78**	**20.20**	**18.65**	**17.37**	**16.43**	**15.77**	**15.47**	**15.43**	**15.67**
NLLLoss-r	26.12	29.69	28.04	27.30	26.30	25.37	24.83	24.27	23.94	23.77	23.76
NLLLoss	**21.63**	**26.45**	**23.90**	**22.46**	**21.36**	**20.44**	**19.88**	**19.46**	**19.30**	**19.25**	**19.41**
TWRLoss-r	**18.97**	**23.02**	**21.11**	**20.01**	**19.03**	**18.13**	**17.40**	**16.85**	**16.65**	**16.58**	**16.81**
TWRLoss	**18.61**	**23.73**	**21.78**	**20.20**	**18.65**	**17.37**	**16.43**	**15.77**	**15.47**	**15.43**	**15.67**
GRU-based-r	**18.64**	**22.79**	**21.29**	**20.05**	**18.82**	**17.64**	**16.79**	**16.25**	**16.02**	**15.90**	**16.16**
GRU-based	20.07	23.73	22.10	21.03	20.04	19.10	18.59	18.19	17.97	18.01	18.27
LSTM-based-r	**18.97**	**23.02**	**21.11**	**20.01**	**19.03**	**18.13**	**17.40**	**16.85**	**16.65**	**16.58**	**16.81**
LSTM-based	**18.61**	**23.73**	**21.78**	**20.20**	**18.65**	**17.37**	**16.43**	**15.77**	**15.47**	**15.43**	**15.67**

Table 3. **Quantitative comparison of different predictors on unseen scenes.** The meanings of VF, TF, and IMU are the same as Table 2.

Comp.	Over. (cm)↓	Early prediction (cm)					Late prediction (cm)				
		10%↓	20%↓	30%↓	40%↓	50%↓	60%↓	70%↓	80%↓	90%↓	100%↓
VF	19.61	24.19	22.56	20.76	19.22	18.34	17.80	17.31	17.02	16.98	17.25
TF+IMU	19.92	23.37	21.94	20.85	19.86	19.11	18.61	18.16	17.87	17.80	17.97
VF+IMU	19.54	23.67	22.07	20.83	19.75	18.70	17.98	17.37	**16.96**	**16.68**	**16.72**
VF+TF	20.82	24.38	23.22	21.93	20.82	19.92	19.28	18.85	18.62	18.59	18.68
VF+TF+IMU	**18.82**	**22.75**	**20.38**	**19.26**	**18.55**	**18.07**	**17.64**	**17.25**	16.97	16.92	17.04
NLLLoss	21.71	25.74	23.63	22.46	21.58	20.89	20.40	19.90	19.60	19.48	19.62
TWRLoss	**18.82**	**22.75**	**20.38**	**19.26**	**18.55**	**18.07**	**17.64**	**17.25**	**16.97**	**16.92**	**17.04**
GRU-based	19.77	23.32	21.81	20.77	19.75	18.96	18.44	17.94	17.67	17.55	17.74
LSTM-based	**18.82**	**22.75**	**20.38**	**19.26**	**18.55**	**18.07**	**17.64**	**17.25**	**16.97**	**16.92**	**17.04**

tures are performed, and the quantitative results are shown in Table 2. We see that: **(1)** the method equipped with all the features achieves the best performance, validating the significance of multimodality features; **(2)** combining visual features, translation, rotation, velocity and acceleration together (VF+TF+IMU) can achieve excellent performance, while missing first-order (VF+IMU) or higher-order (VF+TF) motion features leads to worse performance than the method using visual features only, proving that the first-order and higher-order motion features are both crucial.

Discussions on loss weight. We also compared the performance of the models trained with two different linear loss weight functions: **(1)** linearly reduced from 2 to 1 (stronger penalty at the start), and **(2)** linearly grew from 1 to 2 (stronger penalty at the end). We find that implementing a harsher penalty early on could improve performance.

Discussions on loss function. Compared to Negative Log Likelihood Loss (NLLLoss), our truncated weighted regression loss (TWRLoss) incorporate the grids with higher scores into our training objective, thereby achieving better overall performance in both seen scenes (21.63→18.61) and unseen scenes (21.71→18.82). Meanwhile, the late prediction capability is also improved, for example, in seen scenes, the error could be decreased from 19.41 to 15.67 when observing 100% data.

Table 4. Ablation studies on the granularity of the grid.

Granularity	Over.↓	Early prediction					Late prediction				
		10%↓	20%↓	30%↓	40%↓	50%↓	60%↓	70%↓	80%↓	90%↓	100%↓
$1024^3/m^3$	19.88	24.03	22.08	20.69	19.70	18.86	18.30	17.91	17.71	17.66	17.86
$3072^3/m^3$	19.04	23.36	21.46	20.23	19.15	18.14	17.31	16.64	16.41	16.43	16.73
$5120^3/m^3$	18.61	23.73	21.78	20.20	18.65	17.37	16.43	15.77	15.47	15.43	15.67

Discussions on RNN. LSTM can promote the overall performance a little bit compared to GRU (20.07→18.61 in seen scenes and 19.77→18.82 in novel scenes). This is maybe because LSTM has more learnable parameters than GRU: GRU's bag has two gates (reset and update) while LSTM has three (input, output, forget). A more appropriate architecture design may further improve the prediction.

Ablation study on the grid granularity. The performance gradually improves when the granularity is increased, as shown in Table 4.

5.3. Qualitative Results

We visualize some prediction results in seen scenes and unseen scenes respectively in Fig. 6. We can find that when more temporal information is accumulated, the predictors can generate more precise results. Meanwhile, our baseline approaches demonstrate satisfactory generalization ability. However, there is still a notable gap to be discussed next.

5.4. Limitations and Future works

Dataset. As the initial phase of this academic research, due to limited resources and time, we do not have diverse demographics of skin color, age, weight, height, etc., for our dataset collection participants, nor comprehensive action types. This could limit the generalization ability of models learned on this dataset for any real-world products. However, we believe *the significance of this research is to initiate a useful dataset for relevant academic communities to more easily start to work together on this novel task that*

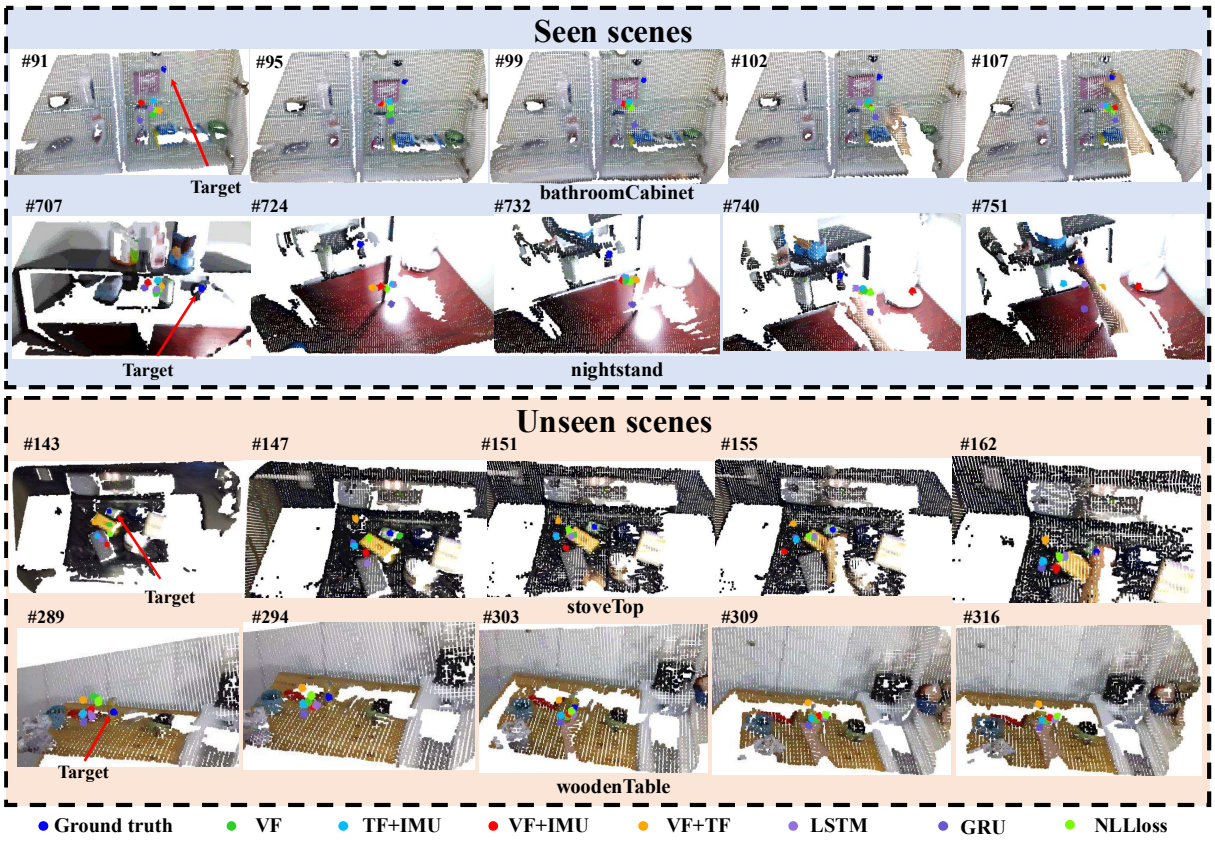

Figure 6. **Qualitative evaluation of different predictors in seen and unseen scenes.** The blue point represents the ground-truth target.

will eventually improve life quality of people with disabilities. Therefore the *initial version of the dataset should not need to be ready for industry usage*. Moreover, as a common practice in datasets of this community, *this paper is not the end of this study* and we are committed to continue growing the dataset to improve the diversity.

Baseline. Our simple RNN baseline achieves a reasonable performance, although not accurate enough: a 20 cm prediction error could still lead to an unintended grasp of a wrong object on the table for a wearable robot user. This could be due to that the multimodality data is only fused with a naive concatenation, which may not be enough to distinguish the importance of different modality at different timestamps. To overcome this limitation, temporal-aware multimodality learning could be a future direction. Moreover, the exploitation of temporal information can also be improved: transformer structure might be more effective than RNN to deal with long sequences. We hope the relevant communities could address these limitations together and improve this task further so as to enable better human-robot collaboration.

Potential negative social impacts. Although our intention of proposing this new task and dataset is to improve human-machine interaction by predicting human intentions, which could help people with disabilities, it is not difficult for cyberpunk Sci-Fi writers to plot evil usages of this to-be-developed technology for building robotic soldiers that are unbeatable by humans. In addition, developing deep learning models has been criticized for its high power consumption and negative impact on climate change.

6. Conclusion

In this work, we propose the first 3D dataset and benchmark for egocentric prediction of action target, which could play a crucial role in wearable devices, human-robot interaction, and augmented reality. Our annotation can be semi-automatic with several off-the-shelf machine learning algorithms, thus is quite efficient. Meanwhile, we design a simple baseline approach based on RNNs to solve the novel task, which is the first method to localize the future action target in 3D. We believe our dataset and benchmark are useful for vision, robotics, and learning communities.

Acknowledgement. The research is supported by NSF FW-HTF program under DUE-2026479. The authors gratefully acknowledge the constructive comments and suggestions from the anonymous reviewers.

References

[1] Khalid El Asnaoui, A. Hamid, A. Brahim, and Ouanan Mohammed. A survey of activity recognition in egocentric lifelogging datasets. In *International Conference on Wireless Technologies, Embedded and Intelligent Systems*, 2017. 2

[2] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *International Conference on Computer Vision*, 2015. 2

[3] A. Bandini and J. Zariffa. Analysis of the hands in egocentric vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[4] Gedas Bertasius, Aaron Chan, and Jianbo Shi. Egocentric basketball motion planning from a single first-person image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Gedas Bertasius, H. Park, Stella X. Yu, and Jianbo Shi. First person action-object detection with egonet. In *Robotics: Science and Systems*, 2017. 2

[6] Minjie Cai, Kris M. Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision*, 2018. 2, 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3

[9] D. Damen, T. Leelasawassuk, Osian Haines, A. Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *British Machine Vision Conference*, 2014. 2

[10] Oriane Dermy, F. Charpillet, and S. Ivaldi. Multi-modal intention prediction with probabilistic movement primitives. In *International Workshop on Human-Friendly Robotics*, 2017. 1

[11] Matteo Dunnhofer, Antonino Furnari, G. Farinella, and C. Micheloni. Is first person vision challenging for object tracking? the trek-100 benchmark dataset. In *International Conference on Computer Vision Workshops*, 2021. 1, 2

[12] Chenyou Fan, Jangwon Lee, and M. Ryoo. Forecasting hands and objects in future frames. In *European Conference on Computer Vision Workshops*, 2018. 2

[13] Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[14] A. Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, 2012. 2, 3

[15] Cornelia Fermüller, Fang Wang, Yezhou Yang, Konstantinos Zampogiannis, Yi Zhang, Francisco Barranco, and Michael Pfeiffer. Prediction of manipulation actions. *International Journal of Computer Vision*, 126(2):358–374, 2018. 3

[16] Antonino Furnari and G. Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision*, 2019. 1, 2

[17] Antonino Furnari and G. Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[18] Harshala Gammulle, S. Denman, S. Sridharan, and C. Fookes. Predicting the future: A jointly learnt model for action anticipation. In *International Conference on Computer Vision*, 2019. 2

[19] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[20] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *International Conference on Computer Vision*, 2021. 2

[21] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[22] Eric Huang, Zhenzhong Jia, and M. T. Mason. Large-scale multi-object rearrangement. In *IEEE International Conference on Robotics and Automation*, 2019. 4

[23] Youngkyoon Jang, Seungtak Noh, H. Chang, Tae-Kyun Kim, and W. Woo. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. *IEEE Trans. Vis. Comput. Graph.*, 21:501–510, 2015. 1

[24] Youngkyoon Jang, Brian T. Sullivan, Casimir J H Ludwig, I. Gilchrist, D. Damen, and W. Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. *International Conference on Computer Vision Workshops*, 2019. 2, 3

[25] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *International Conference on Computer Vision*, 2019. 1, 2

[26] Qiuhong Ke, Mario Fritz, and B. Schiele. Time-conditioned action anticipation in one shot. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[27] Daekyum Kim, B. B. Kang, Kyu Bum Kim, Hyungmin Choi, Jeesoo Ha, Kyu-Jin Cho, and Sungho Jo. Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Science Robotics*, 4, 2019. 2

[28] Alexei A. Efros Krishna Kumar Singh, Kayvon Fatahalian. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 3

[29] Athanasios Krontiris and Kostas E. Bekris. Dealing with difficult instances of object rearrangement. In *Robotics: Science and Systems*, 2015. 4

[30] Yong Jae Lee and Kristen Grauman. Predicting important

objects for egocentric video summarization. *International Journal of Computer Vision*, 114:38–55, 2014. 2

[31] Martin Levihn, Takeo Igarashi, and Mike Stilman. Multi-robot multi-object rearrangement in assignment space. In *IEEE International Conference on Intelligent Robots and Systems*, 2012. 4

[32] Cheng Li and Kris M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *International Conference on Computer Vision*, 2013. 2

[33] Yin Li, Miao Liu, and J. Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3

[34] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[35] Hui Liang, Junsong Yuan, D. Thalmann, and N. Magnenat-Thalmann. Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *ACM Int. Conf. Multimedia*, 2015. 2

[36] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, 2020. 1, 2

[37] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and A. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, 2016. 6

[38] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013. 2

[39] Roberto Martín-Martín, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and S. Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[40] A. Molino, Cheston Tan, Joo-Hwee Lim, and A. Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47:65–76, 2017. 2

[41] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *International Conference on Computer Vision*, pages 8687–8696, 2019. 2

[42] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[43] Eshed OhnBar, Kris Kitani, and Chieko Asakawa. Personalized dynamics models for adaptive assistive navigation systems. In *Conference on Robot Learning*. PMLR, 2018. 2

[44] H. Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3

[45] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *International Conference on Computer Vision*, 2017. 6

[46] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2, 3

[47] F. Ragusa, Antonino Furnari, S. Livatino, and G. Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *IEEE Winter Conference on Applications of Computer Vision*, 2021. 3

[48] Nicholas Rhinehart and Kris M. Kitani. First-person activity forecasting from video with online inverse reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:304–317, 2020. 3

[49] I. Rodin, Antonino Furnari, Dimitrios Mavroedis, and G. Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 2021. 2

[50] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[51] Swathikiran Sudhakaran, S. Escalera, and O. Lanz. Lsta: Long short-term attention for egocentric action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[52] Yansong Tang, Zian Wang, Jiwen Lu, J. Feng, and Jie Zhou. Multi-stream deep neural networks for rgb-d egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:3001–3015, 2019. 1

[53] Tz-Ying Wu, Ting-An Chien, Cheng-Sheng Chan, Chan-Wei Hu, and Min Sun. Anticipating daily intention using on-wrist motion triggered sensing. In *International Conference on Computer Vision*, 2017. 3

[54] Wenxuan Wu, Zhongang Qi, and Fuxin Li. Pointconv: Deep convolutional networks on 3d point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[55] Ryo Yonetani, Kris Kitani, and Yoichi Sato. Visual motif discovery via first-person vision. In *European Conference on Computer Vision*, 2016. 2

[56] Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[57] M. Zhang, K. Ma, J. Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[58] M. Zhang, K. Ma, J. Lim, Qi Zhao, and Jiashi Feng. Anticipating where people will look using adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1783–1796, 2019. 2

[59] Tianyu Zhang, Weiqing Min, Ying-Jie Zhu, Yong Rui, and Shuqiang Jiang. An egocentric action anticipation framework via fusing intuition and analysis. In *ACM Int. Conf. Multimedia*, 2020. 2

[60] Zehua Zhang, David J. Crandall, Chen Yu, and Sven Bambach. From coarse attention to fine-grained gaze: A two-stage 3d fully convolutional network for predicting eye gaze in first person video. In *British Machine Vision Conference*, 2018. 2, 3