# Embodied Human Activity Recognition

Sha Hu Yu Gong Greg Mori
Simon Fraser University
8888 University Drive Burnaby, B.C. Canada. V5A 1S6
hushah@sfu.ca, gongyug@sfu.ca, mori@cs.sfu.ca

## Abstract

*We study how to utilize the mobility of an embodied agent to improve its ability to recognize human activities. We introduce the embodied human activity recognition problem, where an agent moves in a 3D environment to recognize the category of ongoing human activities. The agent must make movement decisions based on its egocentric observations acquired up to the current time, with the goal of choosing movements to obtain new views that lead to accurate human activity recognition. Towards this goal, we propose a reinforcement learning approach that learns a policy controlling the agent's movements over time. We evaluate our approach with two realistic human activity datasets. Results show that our approach can learn to move effectively to achieve high performance in recognizing human activities.*

## 1. Introduction

Building embodied agents that exhibit sensitivity and responsiveness to the presence of humans [1, 46] is a long-standing goal of artificial intelligence. Sensitivity refers to the agents' capability to perceive and understand their surrounding environments - e.g., what humans are doing here. Meanwhile, responsiveness entails the agents' ability to react promptly to their environments.

A key aspect of such human-centered embodied intelligence is that agents cannot afford the luxury of *waiting* for the complete execution of human activities. In many applications, this becomes particularly critical. This includes robotics applications such as assistants in nursing homes, social robots in human environments, or telepresence. For example, in nursing homes, by capturing the subtle signs manifested in individuals' ongoing gaits - such as trembling gaits indicative of body balance loss - these assistive robots can recognize fatal activities, like falls, at an early stage.

In general, an agent could be almost anywhere in a 3D environment. The initial visual observations acquired from the agent's egocentric sensors can be highly unconstrained due to its arbitrary starting positions. These unconstrained visual observations raise challenges in recognizing ongoing human activities. For example, identifying trembling gaits can be very hard by visually sensing from the individual's head level from a top-down view. However, by navigating to observe the individual from the front, the agent can more easily capture the subtle signs of gait progression. Thus, the agent's perceptual sensitivity can benefit from its mobility of moving around to respond proactively. Simultaneously, enhanced perception can guide the agent to plan its movements over time by informing a more accurate understanding of the current progression of human activity.

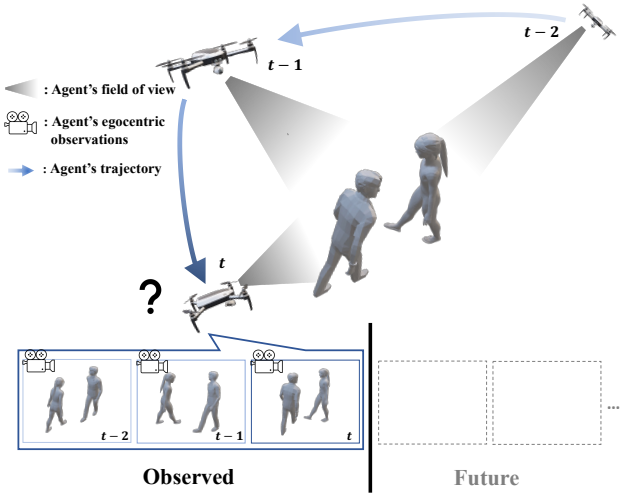



Figure 1. Embodied Human Activity Recognition. An embodied agent is operating in a 3D environment. The agent is tasked to intelligently move around using cues from its egocentric observations so that it can accurately classify an ongoing human activity without seeing the future.

Motivated by these factors, we introduce a novel task: ***Embodied Human Activity Recognition*** (EHAR). In this task, an agent must plan its movements in an environment to recognize the category of an ongoing human activity based on its past and current egocentric observations. As shown in Fig. 1, the human activities occurring in the environment

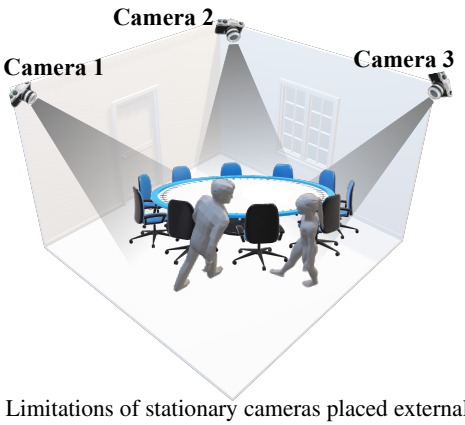



Figure 2. Limitations of stationary cameras placed externally in environments: they require a well-structured environment and cannot proactively respond to dynamic scenarios.

display complex and diverse temporal dynamics. Given that human activities evolve over time, the good viewpoints for observing humans are also expected to change in response to the temporal progression of the activities. This necessitates the agent to continuously reason the activity progression to plan its movements over time strategically.

Our task is related to but different from prior efforts on Human Activity Recognition (HAR) from a third-person sensing perspective. In third-person approaches, sensors such as stationary cameras are placed externally in the environment. As shown in Fig. 2, these cameras are placed so that humans of interest are expected to be centered within their field of view. However, these camera positions are manually pre-determined, requiring prior knowledge of the environment, including the floor plan and furniture geometry. Consequently, these approaches are not directly applicable to unknown and unstructured environments. In addition, stationary cameras cannot proactively respond to dynamic scenarios, particularly when occlusions caused by unexpected objects occur or when good viewpoints change in line with human motions.

Towards addressing the EHAR task, we propose a deep reinforcement learning framework where an agent learns a navigation policy for how to move to early recognize the ongoing human activity. Our key insight is the agent should establish an association between recognition quality changes and its movements. Our proposed agent consists of a recognition model and a policy model, where the latter receives an accumulated recognition state from the former. We evaluate our approach with realistic human activity data with diverse activity classes and complex human motion dynamics. Our agent successfully learns to move around effectively to classify human activity early, outperforming the passive and several heuristic embodied agents. The project page is https://github.com/husha1993/embodied_human_activity_recognition.

## 2. Related Work

**Human Activity Recognition.** As a core problem in computer vision, Human Activity Recognition (HAR) aims to recognize and understand human actions in videos, and has achieved remarkable progress due to deep neural networks. From a data-centric view, early attempts primarily focused on handling RGB data [32]. To capture 3D information and extend the application, recent work pays attention to skeleton [20, 35], point cloud [13] and depth [43] data. These video datasets are captured by stationary cameras, and primarily from a manually selected angle focusing on the humans of interest. However, for EHAR, the robots (or agents) are not static from optimal views. It is required to react to the environment and adjust the camera adaptively from noisy data. From a model-centric view, prior work considered two-stream 2D CNNs [59], RNNs [15], 3D CNNs [16], Transformers [48] and GCNs [12]. These models designed for conventional HAR implicitly assume offline data inputs. However, EHAR is proposed for a more challenging setting of online data streams which necessitates robust and accurate prediction even if the action data are partially observed. The partial observation makes it hard to understand the actions, as the unobserved segments can contain some crucial information. Though *Early Human Activity Recognition* [23, 53] is relevant in the case of recognizing unfinished actions, it still focuses on the data captured by a static camera. On the one hand, it makes EHAR very challenging as the camera view is dynamic, thereby introducing noise and data shifts; on the other hand, it also enables the agents to adaptively react in the environment and avoid accumulated errors from the past states. Therefore, conventional HAR models are not readily applicable to EHAR, and more attention on this task is important.

**Robot-centric Perception of Humans.** In order to expand the habitats of robots from isolated environments to shared human workspaces and social zones, it is essential for robots to understand human behavior through their onboard sensors. Prior work has explored two intertwined aspects of human behavior understanding in the context of robotics: (1) Developing visual sensing of human behaviors tailored for robotic tasks, such as emotion recognition for robot-assisted therapy [38], human trajectory prediction for social robot navigation [39], human intention prediction [54] for human-robot interactions, and human grasp estimation [55] for human-to-robot handovers, etc. (2) Controlling robots to perceive humans within their vicinity, such as drone trajectory optimizations to reconstruct 3D human meshes [41, 47, 60], human tracking in unstructured environments with unpredictable occlusions and motion dynamics [11, 21, 40], human social groups detection via ground robots [61, 62], etc. Our work closely aligns with this line of work as we aim to control an autonomous agent to observe human activities.

Utilizing the opportunity offered by the mobility of robots to yield remarkably flexible human sensing capabilities becomes particularly relevant in light of the rapid growth of various types of commercial affordable robots. Selecting what to see for better perceptual understanding is also related to works on frame or camera selection for HAR.

**Frame Selection or Camera Selection for HAR.** Prior works on frame selection [4, 17, 31, 44, 67–69, 74] aim for efficient and precise human action understanding in long videos by selecting salient frames or clips. However, these approaches either receive a complete action video as input in a passive offline setting [4, 17, 31, 67, 69, 74] or passive streaming setting [44, 68], which are not immediately applicable to an active setting where viewpoints can be manipulated. View selection among multiple cameras [6, 52, 64] aims to select views that offer the best visibility for human action filming or that are most beneficial for human action recognition. However, these approaches are not designed to handle the potentially arbitrary spawning positions and movements of autonomous robots. [26, 58] also tackle sequential viewpoints selections for object recognition, with their focus being on static objects. Consequently, their works do not consider the transient nature of human behaviors in a dynamic real world.

**Human-centered Embodied AI.** In addition to exposing autonomous agents directly to the physical world, embodied AI research provides a complementary paradigm to train and test agents, fueled by the rapid advancements in simulators [2, 37, 56, 70] and the availability of large-scale 3D datasets [3, 71]. However, only a few of these support tasks involve sensing and responding to realistic human activities. [66] learns to control embodied agents to respond to human gestures via a Virtual Reality interface. [50, 51] aims to build assistive agents to interact with virtual humans in household tasks, where these virtual humans are simulated using classical motion planners. The employment of motion planners to simulate virtual humans [36, 45, 73] finds widespread use in both embodied AI tasks and graphic animations.

There are distinctions between the control of embodied agents for robotic tasks and virtual camera control for human motion animations [5, 7, 8, 18, 27, 34]. The primary goal of virtual camera control is to produce storytelling videos of 3D human characters. These studies typically assume full knowledge of the environments, including 4D human motions, while embodied agents can only act on egocentric observations up to the current time. Additionally, a virtual camera agent is not subject to physical constraints, meaning the translation of the camera's position between consecutive time steps can be arbitrarily large. In contrast, the movement strategy for embodied agents needs to account for these constraints of a physical body, which necessitates careful long-horizon planning.

## 3. Task Formulation

We introduce the novel task of Embodied Human Activity Recognition (EHAR). In this task, an autonomous agent starts at an arbitrary location in a 3D environment and is required to recognize an ongoing human activity. The agent observes human activities merely with its onboard sensors. Our goal is to learn a policy that controls the agent to *move around* intelligently to recognize the category of human activity at *each* step of its movement.

**Episode Specification.** An episode is defined by 1) a dynamic human activity scenario $h$ and 2) the agent's starting position $p_0$. The human scenario $h$ is specified by 4D spatial-temporal human skeletons $\boldsymbol{h}^k$ and the activity category $h^c$. $\mathcal{H}$, $\mathcal{H}^c$ and $\mathcal{P}_0$ denote the human scenario space, activity category space, and the agent's starting position space, respectively. In each episode, we sample a human activity scenario $h \sim P(\mathcal{H})$ and an agent's initial position $p_0 \sim P(\mathcal{P}_0)$. In this work, we assume the human motions are within a 3D space of interest. In addition, the center of the space of interest is known a priori and used as the origin of the world coordinate system. The episode length is $L$, meaning the agent is allowed to take up to $L$ actions. Since the task objective is to recognize the activity before fully observing its execution, each task episode begins after the human activity starts and terminates before the activity completes. This is practically simulated by a trimmed activity dataset (with no frames of blank activity) as detailed in Sec. 5.1.

**Sensory Input and Action Space.** The agent makes movement decisions only using its egocentric visual and GPS sensory input. These sensory inputs are not ground-truth human states $h^k$, so they are referred to as partial observations. Concretely, the agent receives egocentric 2D human skeletons $v_t$ and its 3D localization $p_t$ as observation at every step, denoted as $\boldsymbol{o}_t = (v_t, p_t) \in \mathcal{O}$. $v_t$ is represented as a graph with 2D human joint locations as nodes and limbs as edges [12, 72]. The human joint locations are in the pixel coordinate system. In our simulation, $v_t$ is computed by a perspective projection model. For real-world deployment, $v_t$ could be the output of a 2D pose estimator.

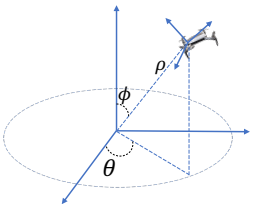


Figure 3. Spherical coordinate system

As shown in Fig. 3, $p_t = (\rho, \theta, \phi)$, where $\rho/\theta/\phi$ represents the agent's distance to the world origin/azimuth/elevation, respectively. $p_t$ is defined on a spherical coordinate system centered on the area of interest, which is a common coordinate system for drone cinematography [10, 22, 25]. This work assumes the agent's visual perception frame always points towards the center of the space of interest, and the visual perception frame and the actuation frame are identical. There-

fore, the action space of the agent reduces to 3 degrees of freedom (DOF). We further discretize the 3-DOF action space into small translation $\Delta_\rho \in \{+0.1m, -0.1m\}$ and rotation $\Delta_\theta, \Delta_\phi \in \{+3.6°, -3.6°\}$. At each time step, the agent chooses to hover or picks one direction of $(\rho, \theta, \phi)$ to move by $\Delta$. Thus, the action space $\mathcal{A}$ consists of $1 + 3 \times 2 = 7$ discrete actions.

**EHAR as a Contextual POMDP.** To highlight the disjoint training/test episodes setting of EHAR, we adopt the Contextual Partially-Observable Markov Decision Process (Contextual POMDP) framework [29], which essentially defines a distribution of POMDPs by introducing a context variable $c$. Formally, the EHAR task space is defined by a tuple $\mathcal{T} = \langle \mathcal{C}, L, \mathcal{S}, \mathcal{O}, \mathcal{A}, T, R \rangle$, where $\mathcal{C}$ denotes the context space. Each task episode $\tau \in \mathcal{T}$ is uniquely specified by a human scenario $h$ and an agent spawning position $p_0$. Thus, the context space $\mathcal{C}$ can be given by $\mathcal{C} = \mathcal{H} \times \mathcal{P}_0$. $\mathcal{S}$ represents the hidden state space, including human motions in 3D space and the ground-truth activity category. $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ represents the state transition probability function. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward function.

**Oracle.** To probe an upper-bound on the performance of an embodied agent, we define an oracle agent that first *imagines* trying all actions exhaustively and receives the ground-truth consequent observations, then makes a hindsight action decision that can yield the lowest cross-entropy loss given the *ground truth* activity label $h^c$, i.e., $a_t = \arg\min_{a \in \mathcal{A}} L_{ce}(h^c, \hat{h}^c_{t+1})$, where $\hat{h}^c_{t+1}$ is the predicted human activity label after receiving the consequent *next time* observation $\boldsymbol{o}_{t+1}$ and $L_{ce}$ is the cross entropy loss. Note that this agent is near-optimal rather than optimal since it acts myopically based on one-step cross-entropy reduction. In addition, it is not realistic in real-world development because it requires knowing the true label and acting with hindsight by peeking into the future.

# 4. Approach

We approach the EHAR task with reinforcement learning. The agent aims to build an incrementally accurate understanding of the ongoing human activities by moving around to gather high-quality observations. Therefore, the movement policy should be informed by the agent's recognition state. Additionally, the agent should be capable of reasoning how the recognition quality changes along with its movements to better plan actions.

Our goal is to learn a policy guided by an activity recognition model. The policy controls the agent movements given the agent's sensory input and accumulated visual comprehension computed by the activity recognition model. We next describe our agent's architecture and its staged training.

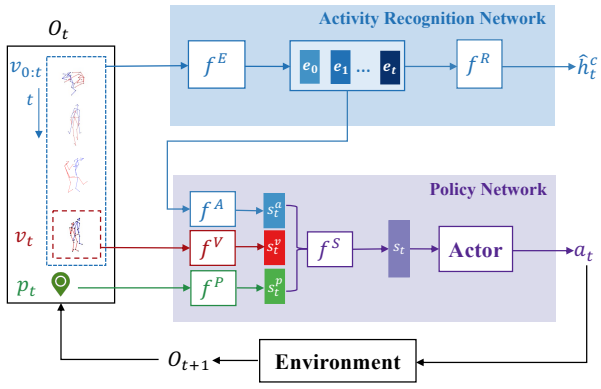



Figure 4. Overview of our agent architecture for EHAR. It has two main components: an activity recognition network and a policy network.

## 4.1. Agent Architecture

As shown in Fig. 4, our agent has two main components: an activity recognition network and a policy network. At each time step $t$, the activity recognition network receives observed human skeletons up to the present $v_{0:t}$ and predicts the activity label $\hat{h}^c_t$. The policy network aims to predict an action $a_t$ executed by the agent. After moving in the 3D space by action $a_t$, the agent acquires a new observation $\boldsymbol{o}_{t+1} = (v_{t+1}, p_{t+1})$, including a new egocentric observation of human skeletons $v_{t+1}$ and the agent's new 3D location $p_{t+1}$. The perception-action loop continues until the agent reaches the task episode horizon $L$.

**Activity Recognition Network.** We adopt the Channel-wise Topology Refinement Graph Convolution Network (CTR-GCN) [12], which achieves start-of-the-art performance on skeleton-based human action recognition. The CTR-GCN consists of a skeleton feature extraction component $f^E$ and a recognition component $f^R$.

**Policy Network.** Movement decisions demand an understanding of the recognition model's state, the current progress of human activities, and the relation between recognition quality and agent position. We encode these three types of information with three separate encoders to inform movement decision-making.

***Accumulated recognition state encoder*** $f^A$. Concretely, the recognition state, obtained from internal representations of the activity recognition network, represents the agent's accumulated comprehension of human activities. We introduce $f^A$ to compute the accumulated recognition state $s^a_t$. It receives visual observations embeddings $e_{0:t}$ extracted from the activity recognition network and outputs $s^a_t$.

***Visual encoder*** $f^V$. It informs the agent of the current progress of human activities. The current visual observation $v_t$ is fed to $f^V$ to compute visual features $s^v_t$.

***Position encoder*** $f^P$. It computes position features $s^p_t$. $s^p_t$ conveys spatial cues, which help the agent to anchor the

relative locations of future high-quality views to its current location.

Those three encoders together enable the agent to reason about how the recognition quality evolves with its movements in 3D space, leading to a better sequential movement policy for long-horizon episodes.

The ***state encoder*** $f^S$ receives a concatenation of the above three states $[s_t^a, s_t^v, s_t^p]$ and outputs a fused state embedding $s_t$. $s_t$ is transformed to a probability distribution over the action space and value of the current state through an actor-critic network.

### 4.2. Agent Training

Following previous works [24, 28, 47], we adopt staged agent training. Namely, we first train the activity recognition network. Then we freeze and plug in the activity recognition network to train the policy network. This staged training is simple yet effective. On the other hand, some prior works [26] also find joint training can lead to a better synergy between perception and action. We leave end-to-end agent training for future work.

**Activity Recognition Network Training.** The activity recognition network is trained by supervised learning using the cross entropy loss. We collect pairs of visual sequences and ground-truth categories at predetermined positions spread out in the 3D space in training episodes. Concretely, we determine 32 positions by a uniform grid of 32 nodes spanning $\mathcal{P}_0$. These positions aim to construct a visual pre-train dataset captured from a wide range of viewpoints. We follow the hyperparameters and optimizers from the prior work [12] to train the activity recognition network.

**Rewards.** The policy is trained with the objective of maximizing accumulative rewards. The policy aims to achieve high overall recognition accuracy over the whole task horizon $L$. For this purpose, We formulate a dense reward $r_t^{acc}$ to encourage the agent to act to maximize accuracy at *every* step. In addition, we want the agent to recognize correctly as early as possible. We therefore define a sparse recognition improvement reward $r_t^{imp}$ that captures the recognition improvement. Typically, the agent outputs incorrect categories in the first few steps, and the correct recognition appears in a certain time step. We want this turning step to happen as early as possible. Towards this goal, the agent should be positively rewarded if it improves its recognition accuracy by moving to produce a better recognition than the *last* timestep. On the other hand, the agent should be penalized if the newly acquired observations lead to worse recognition. Specifically, the reward at each step $r_t$ is given as follows:

$$r_t = \begin{cases} \frac{1}{L} r_t^{acc} & t = 1 \\ \frac{1}{L} r_t^{acc} + r_t^{imp} & 2 \leq t \leq L \end{cases} \quad (1)$$

where $r_t^{acc} = 1$ if $\hat{h}_t^c = h^c$ else 0. $r_t^{imp} = r_t^{acc} - r_{t-1}^{acc}$. The ratio $\frac{1}{L}$ aims to balance $r_t^{acc}$ and $r_t^{imp}$.

**Two-phase Policy Training.** We follow a two-phase training paradigm [9, 42] for policy training. We pre-train the policy with Imitation Learning (IL) [49] and then fine-tune the policy with Proximal Policy Optimization (PPO) [57].

For the IL phase, we first collect a static demonstration dataset from Oracle. It's widely acknowledged [9, 42] that the difficulty of obtaining high accumulated rewards exponentially increases with task horizon. To provide enough demonstrations to overcome the complex optimization landscape of the EHAR task, we collect $10k$ trajectories of oracle performing task episodes randomly sampled from $C_{train}$. We employ the standard behavior cloning objective [49] for IL. For the RL fine-tuning phase, the PPO objective consists of a value network loss, an actor network loss, and an action entropy loss which encourages exploration. See Supp. for details like PPO hyperparameters and network architectures.

## 5. Experiments

To simulate an embodied agent which can move and perceive realistic human activities in 3D space, we use real motion capture data from the Extreme Pose Interaction (ExPI) dataset [19] and the AIST++ dataset [33]. To the best of our knowledge, our work is the first step towards intelligently moving an embodied agent in 3D environments to recognize ongoing human activities. Since there is no existing simulator that can render real human activities within real scenes, we leave experimenting in environments of complex scenes with obstacles and real human activities to future works.

### 5.1. Experimental Setup

**Episode Dataset Preparation.** The episode dataset is constructed as combinations of human scenarios and agent starting positions.

ExPI [19]: ExPI consists of two couples of dancers performing various collective activities. There are 7 classes of common activities performed by different couples of dancers. Each category has 10 mocap sequences, split in 5/5 for training/test set. We further split the training set into training/validation sets in a ratio of 0.85:0.15. The training set is divided into two subsets: one for activity recognition network training and the other for policy training, with the allocation ratio being 0.4:0.6. We take sub-sequences of length $L = 60$ from each sequence to construct episodes with equal temporal horizon, similar to the data preprocessing in the prior work [19]. The starting frame of each sub-sequence is decided by a sliding window of length $L$ and stride $s$. Strides are set to $s = 1/1/60$ for train/val/test, respectively. This results in 5658/998/135 human activity scenarios for training/validation/test sets, i.e., $|\mathcal{H}_{train}| = 5658, |\mathcal{H}_{val}| = 998$, and $|\mathcal{H}_{test}| = 135$.

AIST++ [33]: AIST++ consists of 10 classes of single-

| | | | acc@10 | acc@30 | acc@50 | acc@70 | acc@90 | acc@100 | $\overline{acc}$ |
|---|---|---|---|---|---|---|---|---|---|
| ORACLE | | | 55.56 | 67.41 | 80.74 | 89.63 | 93.33 | 94.81 | 77.53 |
| PASSIVE | NO-ACT | | 49.63 | 47.41 | 51.85 | 57.04 | 61.48 | 62.96 | 52.79 |
| EMBODIED | Heuristic | TOWARDS | 48.89 | 42.96 | 50.37 | 57.78 | 59.26 | 59.26 | 50.64 |
| | | H-ROTATE | 44.44 | 48.15 | 55.56 | 46.67 | 33.33 | 29.63 | 45.23 |
| | | V-ROTATE | 45.19 | 48.15 | 57.04 | 62.22 | 60.74 | 60.00 | 54.52 |
| | | RANDOM | 49.93 ± 1.44 | 45.78 ± 1.22 | 51.26 ± 0.62 | 56.74 ± 2.07 | 60.59 ± 2.74 | 62.07 ± 3.12 | 52.45 ± 0.86 |
| | Learning | PPO ONLY | 46.07 ± 0.81 | 50.52 ± 1.43 | 62.07 ± 3.53 | 67.11 ± 4.40 | 65.63 ± 5.70 | 65.48 ± 6.28 | 58.51 ± 2.45 |
| | | IL ONLY | 49.33 ± 1.12 | 53.04 ± 1.12 | 63.70 ± 2.10 | 65.33 ± 2.53 | 65.78 ± 2.64 | 63.70 ± 3.98 | 59.61 ± 0.78 |
| | | OURS | **50.81** ± 1.24 | **56.30** ± 1.05 | **66.37** ± 1.62 | **71.70** ± 2.06 | **74.81** ± 1.57 | **74.96** ± 1.77 | **64.27** ± 0.33 |

Table 1. Embodied human activity recognition performance on ExPI. Performances of learning-based methods are reported as mean over 5 independent training runs with different seeds.

person activities. Each category has 21 mocap sequences, split in 14/7 for training/test set. We use a process similar to ExPI to sample sub-sequences. See Supp. for more details.

Starting Positions: The agent starting positions $p_0$ are sampled from $\mathcal{P}_0$ uniformly at random, and $\mathcal{P}_0$ is defined as follows: 1) The distance $\rho$ between the initial positions and the world origin ranges from $5m$ to $15m$; 2) The azimuth $\theta$ ranges from $0°$ to $360°$; 3) The elevation $\phi$ ranges from $0°$ to $90°$, meaning that agent is spawned above the ground; and 4) $\mathcal{P}_0$ is a large finite set obtained by a uniform grid spaced by $(\Delta_\rho = 0.1m, \Delta_\theta = 3.6°, \Delta_\phi = 3.6°)$, i.e., $|\mathcal{P}_0| = 100 * 100 * 25 = 250k$.

Task Space: As defined in Sec. 3, each episode is uniquely specified by a human scenario and an agent starting position. The training task space $\mathcal{C}_{train}$ is an exhausted combination of scenario space and position space, e.g., for ExPI, $|\mathcal{C}_{train}| = 5859 * 250k$. val/test episodes are obtained by assigning a single position uniformly sampled from $\mathcal{P}_0$ to each scenario, e.g., $|\mathcal{C}_{val}| = 140$ and $|\mathcal{C}_{test}| = 135$ for ExPI.

**Evaluation Metrics.** We use standard metrics [53, 64]: 1) accuracy at an observation ratio of $x\%$ denoted as $acc@x$, where $x \in \{10, 30, 50, 70, 90, 100\}$; and 2) the average accuracy across all time steps, denoted as $\overline{acc}$. $\overline{acc}$ measures the overall *early* recognition accuracy over different observation ratios. We also plot the accuracy against observation ratios as a curve.

**Baselines.** We compare against the following methods:

- `No-Act`: an agent takes no movement action and holds its starting position for all steps, representing a passive policy.
- `Random`: an agent randomly selects an action from the action space $\mathcal{A}$.
- `Towards`: an agent moves towards the humans so that it is likely to recognize the activity better.
- `H-Rotate` [75]: an agent undergoes a constant horizontal rotation to observe human activity from a wide range of viewpoints.
- `V-Rotate`: an agent always decreases its elevations to move towards an identical horizontal level with humans.

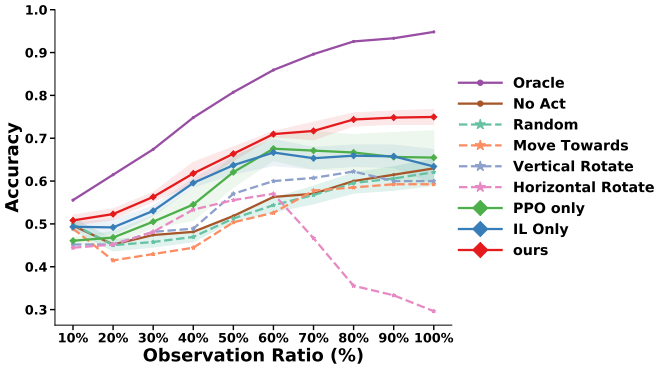



Figure 5. Curve of accuracy against observation ratios on ExPI. Our full model shows increasing accuracy as acquiring more observations. This suggests that our model can learn more effective movement behaviors than other baselines.

## 5.2. Results and Findings

**Our Model Outperforms Baselines across Observation Ratios.** Table. 1 and Table. 2 show the accuracy results across observation ratios and the average accuracy on ExPI and AIST++, respectively. Passive agents perform poorly compared to embodied agents that can move. This proves that embodied agents capable of navigating within a 3D environment can better perceive human activity.

All variants of learning-based methods outperform the heuristic baselines in overall early recognition accuracy $\overline{acc}$, showing that our proposed model can learn an intelligent policy that can more effectively control an agent in gathering high-quality views.

We show the curve of accuracy against observation ratios on ExPI in Fig. 5. Our full model outperforms others across different observation ratios. In addition, our model shows an increasing trend in accuracy as more observations are acquired. This means our model can better leverage the observations acquired up to the current timestep and strategically move around for enhanced recognition in subsequent timesteps. In contrast, `H-Rotate` on ExPI shows a significant drop in accuracy when acquiring more observations. By examining the sequential observations of `H-Rotate`, we find that the egocentric human skeletons in pixel space

| | | | acc@10 | acc@30 | acc@50 | acc@70 | acc@90 | acc@100 | $\overline{\text{acc}}$ |
|---|---|---|---|---|---|---|---|---|---|
| ORACLE | | | 21.24 | 40.71 | 57.52 | 60.18 | 69.91 | 69.03 | 51.22 |
| PASSIVE | | NO-ACT | 20.35 | 30.97 | 37.17 | 41.59 | 40.71 | 43.36 | 32.68 |
| EMBODIED | Heuristic | TOWARDS | 19.47 | 29.20 | 36.28 | 35.39 | 39.82 | 42.48 | 31.86 |
| | | H-ROTATE | 19.47 | 27.43 | 35.39 | 39.82 | 40.70 | 44.24 | 32.77 |
| | | V-ROTATE | 16.81 | 23.01 | 33.63 | 34.51 | 37.17 | 39.82 | 28.08 |
| | | RANDOM | 18.89 ± 1.02 | 27.14 ± 0.51 | 36.87 ± 0.51 | 38.94 ± 0.51 | 37.61 ± 3.13 | 41.15 ± 3.13 | 32.06 ± 1.09 |
| | Learning | PPO ONLY | 17.88 ± 1.58 | 35.04 ± 3.10 | 46.90 ± 2.34 | 53.81 ± 3.62 | 58.58 ± 3.51 | 59.47 ± 3.88 | 42.59 ± 1.57 |
| | | IL ONLY | 17.26 ± 0.51 | 31.42 ± 2.11 | 42.04 ± 1.53 | 45.58 ± 3.27 | 45.13 ± 1.91 | 43.58 ± 2.64 | 36.38 ± 0.63 |
| | | OURS | 19.25 ± 1.96 | **37.61** ± 1.14 | **50.00** ± 2.34 | **58.41** ± 2.17 | **63.72** ± 2.29 | **67.04** ± 3.34 | **46.06** ± 0.50 |

Table 2. Embodied human activity recognition performance on AIST++. Performances of learning-based methods are reported as mean over 5 independent training runs with different seeds.

undergo substantial transformations over time as the agent constantly changes its azimuths within an episode. These observation transformations include both 2D joints positions of each person and relative positions of person to person. This temporal incoherence in the pixel space between consecutive frames within one episode, resulted by changes in viewpoints, is known as *shotcuts* [14], which can adversely impact recognition accuracy. More observations do not necessarily lead to improved recognition. We also notice that `H-Rotate` on AIST++ does not show decreases in $acc@x$ when $x$ increases in Table. 2. Since activities from AIST++ only involve a single person, the observation transformations of `H-Rotate` do not include relative positions between humans, resulting in less recognition confusion.

**EHAR is a Hard-Exploration Problem and IL helps.** To better understand the challenges posed by the EHAR task, we ablate various training stages of our learned agents in Table. 1 and in Table. 2.

The PPO-ONLY agent, which learns an intelligent moving policy better than other heuristic baselines, shows that manual reward engineering and action entropy regularization (Sec. 4.2) can to some extent mitigate the hard exploration. However, we empirically find that the stability of PPO training curves depends on the network initialization and action sampling, both of which are determined by experimental seeds. Moreover, the learned policies are prone to converge to a naive behavior mode, wherein a specific action consistently has a high probability throughout an episode. Policy initialization and sampled actions heavily influence the initial behavior trajectories during training. These trajectories subsequently impact the initial policy optimization, often resulting in a sub-optimal strategy. Thus, we conjecture that the unstable and sub-optimal behaviors of PPO-ONLY agents stem from an inefficient exploration of the environments at the beginning of policy learning, and they struggle to recover in later interactive training.

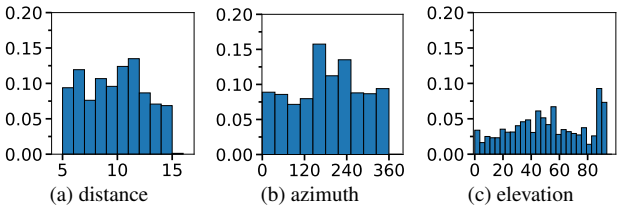
The IL-ONLY agent, which learns a policy through supervised learning on a static demonstration dataset, shows more stable learning than PPO from scratch. In addition, the IL agent converges to diverse modes of action selection behavior instead of a single mode. Thus, we hypothesize that IL helps with the hard-exploration issue of EHAR. Furthermore, by fine-tuning the IL pretrained agent with PPO, the agent obtains further performance improvement. We hypothesize that the improvement of OURS over IL-ONLY results from mitigating the difficulty of imitating an oracle that requires privileged information [65]. Such privileged information, including hindsight action selection by peeking into the future and ground truth activity label, is not available as observations to the agent during training. The interactive training with environments through PPO fine-tuning can bypass the imitation difficulty of privilege absence.

| | acc@10 | acc@50 | acc@100 | $\overline{\text{acc}}$ |
|---|---|---|---|---|
| w/o $f^A$ | 50.93 ± 1.64 | 64.81 ± 0.96 | 67.59 ± 2.86 | 61.32 ± 0.80 |
| w/o $f^V$ | 48.33 ± 1.64 | 62.96 ± 1.35 | 71.30 ± 2.52 | 61.41 ± 0.86 |
| w/o $f^P$ | 52.35 ± 0.43 | 66.91 ± 0.43 | 70.62 ± 3.50 | 63.36 ± 0.81 |
| ours | 50.81 ± 1.24 | 66.37 ± 1.62 | 74.96 ± 1.77 | 64.27 ± 0.33 |

Table 3. Ablation of policy components of our agent on ExPI.

**Ablations of Policy Components.** To study the different roles of policy inputs for EHAR, we report the ablation of policy input components on ExPI in Table. 3. The accumulated recognition state encoder $f^A$ plays an important role in the overall performance. This suggests the importance of encoding accumulated visual understanding over a sequence of visual observations acquired up to the present. Given that $f^V$ is learned with visual observations acquired by a policy online, it can encode visual features that match with the policy behavior, accounting for the view transformations along with movements during both training and testing phases. The pose encoder $f^P$ is also critical since it informs the agent of its spatial positions, building an implicit association of recognition quality and its current 3D position over time.

**Behavior Mode of the Learned Agent.** We provide statistics and analysis on the behavior mode of the learned agent on ExPI. Fig. 6 and Fig. 7a-7d show the agents' temporal and spatial behavior patterns, respectively. As shown in Fig. 6, the spatial positions of the agent exhibit a broad distribution across various elevations $\phi$, distance to the world origin $\rho$, and azimuths $\theta$. This implies that the agent's navigational trajectories span widely across 3D space, indicating that the agent learns to actively adjust its positions within its




(a) distance (b) azimuth (c) elevation

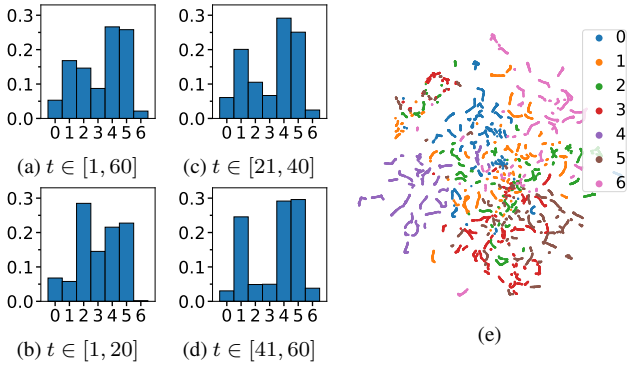
Figure 6. Histograms of distance (m), azimuth (degree), and elevation of agent positions (degree).

environment. In addition, the distributions over the three dimensions are nearly uniform. It means that there is no canonical position that is especially favored for EHAR; instead, high-quality views are acquired from diverse positions. Finally, the peaks at elevations between ranges of 85°and 90°(i.e., $p(\phi \in [85°, 90°]) \approx 0.2$ in Fig. 6c suggest that the learned agent tends to move to positions at a same horizontal level with humans. This tendency can explain the relatively good performance of `V-Rotate` among other heuristic agents.

Fig. 7a shows the probability distribution of the agent's actions computed from 135 test episodes. We further investigate the distribution of the agent's actions across different stages of episodes. We segment the full episode into three stages of equal timesteps, i.e., $t \in [1, 20]$ in Fig. 7b, $t \in [21, 40]$ in Fig. 7c, $t \in [41, 60]$ in Fig. 7d. This helps to understand the evolution of the agent's behavior patterns over time. The action indexes are defined as follows: $a = 0/1$ means move to increase/decrease $\rho$ by $\Delta_\rho$; $a = 2/3$ means move to increase/decrease $\theta$ by $\Delta_\theta$; $a = 4/5$ means move to increase/decrease $\phi$ by $\Delta_\phi$; $a = 6$ means hovering.

We notice that the probability of hovering (i.e., $a_t = 6$) tends to increase as the episodes progress. This trend might result from the agent's active adjustments of its 3D positions in the initial stages of the episodes, followed by its tendency to keep static after moving to some good viewpoints. However, the probability of taking other actions to move around (i.e., $a_t \neq 6$) is still high in the later stage, indicating that the good viewpoints keep changing. Since human activity is dynamic through time, the good viewpoints for observing humans are also expected to change over time along with the temporal progression of human motions.

**Visualizations of Learned State Representations.** We examine the state representations $s_t$, which are the outputs of the state encoder $f^S$ at each timestep of episodes, to understand the learned agent's successful performance in the EHAR task. Fig. 7e shows the t-SNE [63] visualization of $s_t$ on ExPI. Each point, denoting a state representation, is colored by the ground truth human activity label of the corresponding episode. We find that the clusters can correspond well to the human activity classes. Given that the agent's actions are directly mapped from the state representations, we conclude that the agent learns to move by considering the activity semantics. This means that our agent builds a tight synergy between its perception and action.




(a) $t \in [1, 60]$ (c) $t \in [21, 40]$

(b) $t \in [1, 20]$ (d) $t \in [41, 60]$ (e)

Figure 7. (a), (b), (c), (d) represent histograms of the agent's actions across different stages of episodes. (e) is the t-SNE visualization of learned state embeddings.

**Predictability of Human Activity for Embodied Perception.** The *predictability* of human activity varies across activity types and perception settings. We show that the predictability of human activity in the embodied setting is different from the passive setting. Specifically, the predictability of human activity is defined by the portion of an activity sequence that needs to be observed before being classified correctly [20, 30]. Following prior works, three categories of predictability are defined: *Instantly preditable* (IP), *Early preditable* (EP), and *Late preditable* (LP). IP/EP/LP means the activity sequence can be correctly classified after observing 10%/50%/100%, respectively. As shown in Table. 4, though the number of activity types that are instantly or early predictable is larger for embodied perception than for passive perception, it's still challenging to recognize most activity types given less than 50% observations in both settings.

| | Instantly Predictable | Early Predictable | Late Predictable |
|---|---|---|---|
| **Passive Perception** | noser | cartwheel; rog-classic | around-the-back; coochie; a-frame; toss-out |
| **Embodied Perception** | noser; rog-classic | toss-out; coochie; cartwheel | a-frame; around-the-back; |

Table 4. Embodied predictability and passive predictability of different human activities on ExPI.

## 6. Conclusion

In this work, We introduce the EHAR task – an agent is spawned in a 3D environment and is able to move in order to recognize ongoing human activities by acquiring high-quality observations. To tackle this task, we propose a reinforcement learning approach to learn an intelligent movement policy. Through quantitative comparisons with various baselines, we demonstrate the importance of strategic movements for EHAR. In addition, through several ablation experiments and qualitative analysis, we show that our proposed agent can learn effective movement behavior to achieve high performance in recognizing human activities.

## References


[1] *The Artificial life route to artificial intelligence : building embodied, situated agents / edited by Luc Steels and Rodney Brooks.* L. Erlbaum Associates, 1994. 1

[2] AI2-THOR: An Interactive 3D Environment for Visual AI. *ArXiv*, abs/1712.05474, 2017. 3

[3] ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. *NeurIPS 2022*, 2022. 3

[4] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 251–266, 2018. 3

[5] Rao Anyi, Jiang Xuekun, Guo Yuwei, Xu Linning, Yang Lei, Jin Libiao, Lin Dahua, and Dai Bo. Dynamic storyboard generation in an engine-based virtual environment for video production. *arXiv preprint arXiv:2301.12688*, 2023. 3

[6] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 3

[7] Jackie Assa, Yaron Caspi, and Daniel Cohen-Or. Action synopsis: pose selection and illustration. *ACM Transactions on Graphics (TOG)*, 24(3):667–676, 2005. 3

[8] Jackie Assa, Daniel Cohen-Or, I-Cheng Yeh, and Tong-Yee Lee. Motion overview of human actions. *ACM Transactions on Graphics (TOG)*, 27(5):1–10, 2008. 3

[9] Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando De Freitas. Playing hard exploration games by watching youtube. *Advances in neural information processing systems*, 31, 2018. 5

[10] Rogerio Bonatti, Arthur Bucker, Sebastian Scherer, Mustafa Mukadam, and Jessica Hodgins. Batteries, camera, action! learning a semantic control space for expressive robot cinematography. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7302–7308. IEEE, 2021. 3

[11] Rogerio Bonatti, Cherie Ho, Wenshan Wang, Sanjiban Choudhury, and Sebastian Scherer. Towards a robust aerial cinematography platform: Localizing and tracking moving targets in unstructured environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 229–236. IEEE, 2019. 2

[12] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 2, 3, 4, 5

[13] Huaining Cheng and Soon Myoung Chung. Orthogonal moment-based descriptors for pose shape query on 3d point cloud patches. *Pattern Recognit*, 2016. 2

[14] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 269–284. Springer, 2016. 7

[15] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[16] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2

[17] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 3

[18] Christoph Gebhardt and Otmar Hilliges. Optimization-based user support for cinematographic quadrotor camera target framing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021. 3

[19] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 5

[20] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, 129(7):2097–2112, 2021. 2, 8

[21] Benjamin Hepp, Tobias Nägeli, and Otmar Hilliges. Omnidirectional person tracking on a flying robot using occlusion-robust ultra-wideband signals. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 189–194, 2016. 2

[22] Cherie Ho, Andrew Jong, Harry Freeman, Rohan Rao, Rogerio Bonatti, and Sebastian Scherer. 3d human reconstruction in the wild with collaborative aerial cameras. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5263–5269. IEEE, 2021. 3

[23] Minh Hoai and Fernando De la Torre. Max-margin early event detectors. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2863–2870, 2012. 2

[24] Chong Huang, Fei Gao, Jie Pan, Zhenyu Yang, Weihao Qiu, Peng Chen, Xin Yang, Shaojie Shen, and Kwang-Ting Cheng. Act: An autonomous drone cinematography system for action scenes. In *2018 ieee international conference on robotics and automation (icra)*, pages 7039–7046. IEEE, 2018. 5

[25] Chong Huang, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, and Kwang-Ting Cheng. Through-the-lens drone filming. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4692–4699, 2018. 3

[26] Dinesh Jayaraman and Kristen Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 489–505. Springer, 2016. 3, 5

[27] Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. Example-driven virtual cinematography by learn-

ing camera behaviors. *ACM Transactions on Graphics (TOG)*, 39(4):45–1, 2020. 3

[28] Sena Kiciroglu, Helge Rhodin, Sudipta N Sinha, Mathieu Salzmann, and Pascal Fua. Activemocap: Optimized viewpoint selection for active human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2020. 5

[29] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, jan 2023. 4

[30] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1481, 2017. 8

[31] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019. 3

[32] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *International Conference on Computer Vision (ICCV)*, 2011. 2

[33] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 5

[34] Christophe Lino and Marc Christie. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)*, 34(4):1–12, 2015. 3

[35] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding, 2017. 2

[36] Wenhan Luo, Peng Sun, Fangwei Zhong, Wei Liu, Tong Zhang, and Yizhou Wang. End-to-end active object tracking via reinforcement learning. In *International conference on machine learning*, pages 3286–3295. PMLR, 2018. 3

[37] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 3

[38] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian Sminchisescu. 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2018. 2

[39] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *J. Hum.-Robot Interact.*, 12(3), apr 2023. 2

[40] Sepehr MohaimenianPour and Richard Vaughan. Hands and faces, fast: Mono-camera user detection robust enough to directly control a uav in flight. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5224–5231, 2018. 2

[41] Tobias Nägeli, Samuel Oberholzer, Silvan Plüss, Javier Alonso-Mora, and Otmar Hilliges. Flycon: Real-time environment-independent multi-view human pose estimation with aerial vehicles. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018. 2

[42] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018. 5

[43] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *International Conference on Computer Vision Workshops (ICCVW)*, 2011. 2

[44] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7576–7585, 2021. 3

[45] Claudia Pérez-D'Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. Robot navigation in constrained pedestrian environments using reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1140–1146. IEEE, 2021. 3

[46] Rolf Pfeifer and Fumiya Iida. Embodied artificial intelligence: Trends and challenges. In *Embodied artificial intelligence*, pages 1–26. Springer, 2004. 1

[47] Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 5

[48] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 2021. 2

[49] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988. 5

[50] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 3

[51] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-{ai} collaboration. In *International Conference on Learning Representations*, 2021. 3

[52] Dmitry Rudoy and Lihi Zelnik-Manor. Viewpoint selection for human actions. *International journal of computer vision*, 97:243–254, 2012. 3

[53] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 international conference on computer vision*, pages 1036–1043. IEEE, 2011. 2, 6

[54] Michael S Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 295–302, 2015. 2

[55] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 51–69. Springer, 2022. 2

[56] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 3

[57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[58] Soroush Seifi, Abhishek Jha, and Tinne Tuytelaars. Glimpse-attend-and-explore: Self-attention for active visual exploration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16137–16146, 2021. 3

[59] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 2014. 2

[60] Rahul Tallamraju, Nitin Saini, Elia Bonetto, Michael Pabst, Yu Tang Liu, Michael J Black, and Aamir Ahmad. Aircaprl: autonomous aerial human motion capture using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4):6678–6685, 2020. 2

[61] Angelique Taylor, Darren M Chan, and Laurel D Riek. Robot-centric perception of human groups. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(3):1–21, 2020. 2

[62] Angelique Taylor and Laurel D Riek. Regroup: A robot-centric group detection and tracking system. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 412–421. IEEE, 2022. 2

[63] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[64] Boyu Wang, Lihan Huang, and Minh Hoai. Active vision for early recognition of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1081–1091, 2020. 3, 6

[65] Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. Bridging the imitation gap by adaptive insubordination. *Advances in Neural Information Processing Systems*, 34:19134–19146, 2021. 7

[66] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4095–4102. IEEE, 2021. 3

[67] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6222–6231, 2019. 3

[68] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. *Advances in neural information processing systems*, 32, 2019. 3

[69] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 3

[70] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018. 3

[71] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022. 3

[72] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[73] Ruolin Ye, Wenqiang Xu, Haoyuan Fu, Rajat Kumar Jenamani, Vy Nguyen, Cewu Lu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Rcareworld: A human-centric simulation world for caregiving robots. 2022.

[74] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2678–2687, 2016. 3

[75] Xiaowei Zhou, Sikang Liu, Georgios Pavlakos, Vijay Kumar, and Kostas Daniilidis. Human motion capture using a drone. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2027–2033. IEEE, 2018. 6