Egocentric Activity Recognition and Localization on a 3D Map

Miao Liu[1]⋆ , Lingni Ma[5], Kiran Somasundaram[5], Yin Li[2], Kristen Grauman[3,4], James M. Rehg[1], Chao Li[5]

[1] Georgia Institute of Technology
[2] University of Wisconsin-Madison
[3] The University of Texas at Austin
[4] Meta AI
[5] Meta Reality Labs

Abstract. Given a video captured from a first person perspective and the environment context of where the video is recorded, can we recognize what the person is doing and identify where the action occurs in the 3D space? We address this challenging problem of jointly recognizing and localizing actions of a mobile user on a known 3D map from egocentric videos. To this end, we propose a novel deep probabilistic model. Our model takes the inputs of a Hierarchical Volumetric Representation (HVR) of the 3D environment and an egocentric video, infers the 3D action location as a latent variable, and recognizes the action based on the video and contextual cues surrounding its potential locations. To evaluate our model, we conduct extensive experiments on the subset of Ego4D dataset, in which both human naturalistic actions and photo-realistic 3D environment reconstructions are captured. Our method demonstrates strong results on both action recognition and 3D action localization across seen and unseen environments. We believe our work points to an exciting research direction in the intersection of egocentric vision, and 3D scene understanding.

Keywords: Egocentric Vision, Activity Recognition, 3D Scene Understanding

1 Introduction

Egocentric vision has emerged as a promising paradigm for understanding human activities in a mobile setting. Its defining characteristic is the continuous capture of first-person visual experience. In particular, egocentric videos implicitly and naturally connect the camera wearer's activities to the relevant 3D spatial context, such as the surrounding objects and their 3D layout. While this observation has been true since the beginning of egocentric vision, it is only recently that 3D scene models that can capture this context have become readily available, due to advances in 3D scanners [57] and Augmented Reality (AR) headsets [22]. Fig. 1

⋆ This work was primarily done during an internship at Meta Reality Labs.
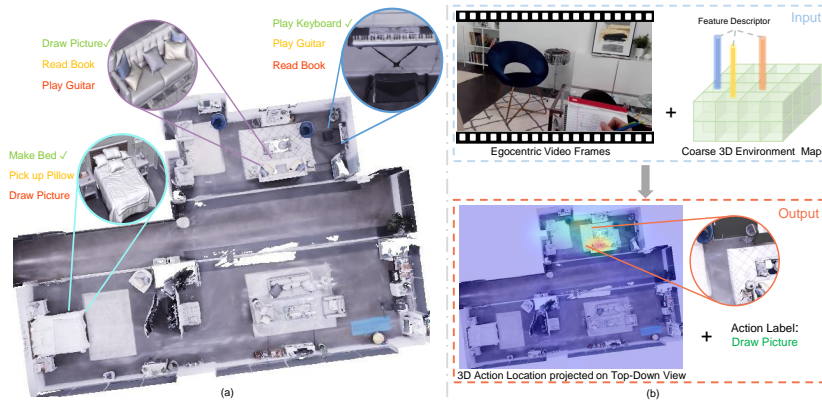
Fig. 1. (a) The activities of daily life take place in a 3D environment, and the semantic and spatial properties of the environment are powerful cues for activity recognition. (b) *Our Proposed Task*: Given an input egocentric video sequence and a 3D volumetric representation of the environment (carrying both semantic and geometric information), our goal is to detect and localize activities, by jointly predicting the action label and location on the 3D map where it occurred.

(a) gives an example of a 3D scan of a subject's apartment in which the 3D layout of the furniture and appliances is known *a priori*. Given an egocentric video, our goal is to leverage this 3D map to reason about the camera wearer's activities and the 3D locations in which they are performed, *e.g.* drawing a picture while sitting on the sofa. Such a capability could enable future context-sensitive applications in AR and Human-Robot Interaction (HRI).

This paper addresses the following research question: *How can we design vision models to exploit the prior knowledge of a known 3D environment for recognizing and localizing egocentric activities?* This question has not been tackled by existing works on egocentric action and activity recognition [23, 67, 34, 31, 41, 36, 43, 29]. Prior works have used limited contextual cues for egocentric video analysis, such as a 2D ground plane [46] or a topological map [38], and have focused on understanding the functions of an environment, such as the common locations at which activities occur. In contrast, this work introduces the new task of the *joint recognition and 3D localization of egocentric activities given trimmed videos and a coarsely-annotated 3D environment map*. We provide a visual illustration of our problem setting in Fig. 1 (b).

Two major challenges arise in our task. First, standard architectures for egocentric activity recognition are not designed to incorporate 3D scene context, requiring a new design of action recognition models and associated 3D scene representations. Second, the exact ground truth for the locations of actions in a 3D scene that is the size of an entire apartment is difficult to obtain, due to ambiguities in 2D to 3D registration. As a remedy, we leverage camera registration using structure-from-motion that yields "noisy" locations, which requires the model to address the uncertainty in action locations during training.

To address the challenge of leveraging context in recognition, we develop a Hierarchical Volumetric Representation (HVR) to describe the semantic and geometric information of the 3D environment map (see Fig. 2 (a) and Sec 3.1 for explanation). We further present a novel deep model that takes egocentric videos and our proposed 3D environment HVR as inputs, and outputs the 3D action locations and the activity classes. Our model consists of two branches. The *environment branch* makes use of a 3D convolutional network to extract global environmental features from HVR. Similarly, the *video branch* uses a 3D convolutional network to extract visual features from the input video. The environmental and visual features are further combined to estimate the 3D activity location, supervised by the results of camera registration. Moreover, we tackle the second challenge of noisy localization by using stochastic units to account for uncertainty. The predicted 3D activity location, in the form of a probabilistic distribution, is then used as a 3D attention map to select local environmental features relevant to the action. Finally, these local features are further fused with video features for recognition.

Our method is trained and evaluated on the recent, freely-available Ego4D dataset [13], which contains naturalistic egocentric videos and photo-realistic 3D scene reconstructions along with 3D static object annotations. We demonstrate strong results on action recognition and 3D action localization. Specifically, our model outperforms a strong baseline of 2D video-based action recognition methods by 4.2% in mean class accuracy, and beats baselines on 3D action localization by 9.3% in F1 score. Furthermore, we demonstrate that our method can generalize to unseen environments not present in the training set yet with known 3D maps and object labels. We believe this work provides a useful foundation for egocentric video understanding in a 3D scene context

2   Related Work

We first discuss the most relevant works on egocentric vision, and then review several previous efforts on human-scene interaction and 3D scene representation. **Egocentric Vision**. There is a rich set of literature aiming at egocentric activity understanding. Prior works have made great progress in recognizing and anticipating egocentric actions based on 2D videos [10, 52, 24, 67, 34, 31, 41, 36, 43], and predicting gaze and locomotion [29, 28, 18, 55, 39, 64, 40, 42, 51]. Far fewer works have considered environmental factors and spatial grounding of egocentric activity. Guan et al. [14] and Rhinehart et al. [47] jointly considered trajectory forecasting and egocentric activity anticipation with online inverse reinforcement learning. The most relevant works to ours are recent efforts on learning affordances for egocentric action understanding [38, 46]. Nagarajan et al. [38] introduced a topological map environment representation for long-term activity forecasting and affordance prediction. Rhinehart et al. [46] considered a novel problem of learning "Action Maps" from egocentric videos. However, methods that use ground plane representations of the environment [46] or environmental functionality as the context [38] may lack the specificity provided by 3D prox-

imity. In contrast to these prior efforts, our focus is on exploiting the geometric and semantic information in the HVR map to address our novel task of joint egocentric action recognition and 3D localization.

Human-Scene Interaction. Human-scene constraints have been proven to be effective in estimating human body model [66, 16, 65]. The most relevant prior works focus on understanding environment affordance. Grabner et al. [12] predict object functionality by hallucinating an actor interacting with the scene. A similar idea was also explored in [20, 21].Koppula et al. [25] leveraged RGB-D videos to jointly recognize human activities and estimate objects affordances. Savva et al. [49] predicted action heat maps that highlight the likelihood of an action in the scene by partitioning 3D scenes into disjoint sets of segments and learning a segment dictionary. Gupta et al. [15] presented a human-centric scene representation for predicting the afforded human body poses. Delaitre et al. [5, 8] introduced a statistical descriptor of person-object interactions for object recognition and human body pose prediction. Fang et al. [7] proposed to learn object affordances from demonstrative videos. Nagarajan et al. [37] proposed to use backward attention to approximate the interaction hotspots of future action. Those previous efforts were limited to the analysis of environment functionality [20, 21, 5, 8], constrained human action and body pose [49], or hand-object interaction on 2D image plane [7, 37]. In contrast, we are the first to utilize the rich geometric and semantic information of the 3D scene for naturalistic human activity recognition and 3D localization.

3D Scene Representation. Many recent works explored various 3D representations for 3D vision tasks, including 3D object detection [63, 68, 3] and embodied visual navigation [11, 17, 27]. Deep models have been developed for point clouds [44, 61, 60, 45] with great success in object recognition, semantic segmentation, and sceneflow estimation. However, using point clouds to describe a large-scale 3D scene will result in high computational and memory cost [68]. To address this challenge, many approaches used rasterized point clouds in a 3D voxel grid, with each voxel represented by either handcrafted features [58, 6, 53, 54] or learning-based features [68]. Signed-Distanced value and Chamfer Distance between 3D scene and 3D human body have also been used to enforce more plausible human-scene contact [66, 16, 65, 33]. Building on this prior works, we utilize a 3D Hierarchical Volumetric Representation (HVR) that encodes the geometric and semantic context of a 3D scene for egocentric activity understanding.

3   Method

We denote a trimmed input egocentric video as $x = (x^1, ..., x^t)$ with frames x^t indexed by time t. In addition, we assume a global 3D environment prior e, associated each with input video, is available at both training and inference time. e is environment specific, *e.g.* the 3D map of an apartment. Our goal is to jointly predict the action category y of x and the action location r on the 3D map. r is parameterized as a 3D saliency map, where the value of $r(w, d, h)$ represents the likelihood of action clip x happening in spatial location w, d, h. For tractability,
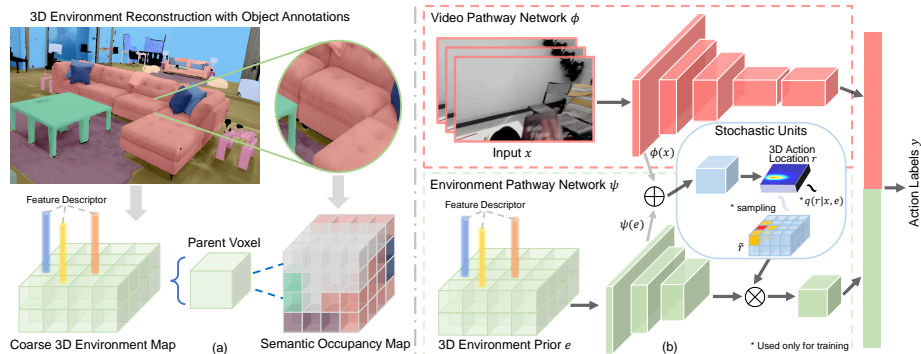
Fig. 2. (a) Hierarchical Volumetric Representation (HVR). We rasterize the semantic 3D environment mesh into two levels of 3D voxels. Each parent voxel corresponds to a possible action location, while the children voxels compose a semantic occupancy map that describes their parent voxel. (b) Overview of our model. Our model takes video clips x and the associated 3D environment representation e as inputs. We adopt an I3D backbone network ϕ to extract video features and a 3D convolutional network ψ to extract the global environment features. We then make use of stochastic units to generate sampled action location $\tilde{r}$ for selecting local 3D environment features for action recognition. Note that $\otimes$ represents weighted average pooling, while $\oplus$ denotes concatenation along channel dimension.

we associate the entire activity with a specific 3D location and do not model location change over the course of an activity. This is a valid assumption for the activities we address, such as sitting down, playing keyboards, etc. r thereby defines a proper probabilistic distribution in 3D space.

In this section, we first introduce our proposed joint model of action recognition and 3D localization, leveraging a 3D representation of the environment. We then describe key components of our model, training and inference schema, as well as our network architecture.

3.1 Joint Modeling with the 3D Environment Representation

3D Environment Representation. We seek to design a representation that not only encodes the 3D geometric and semantic information of the 3D environment, but is also effective for 3D action localization and recognition.

To this end, we introduce a Hierarchical Volumetric Representation (HVR) of the 3D environment. We provide an illustration of our method in Fig. 2(a). We assume the 3D environment reconstruction with object labels is given in advance as a 3D mesh (see Sec.4 for details). We first divide the 3D mesh into $X \times Y \times Z$ parent voxels, that define all possible action locations. We then divide each parent voxel into multiple voxels at a fixed resolution M and further assign an object label to each child voxel based on the object annotation. Specifically, the object label of each child voxel is determined by the majority vote of the vertices

that lie inside that child voxel. Note that we only consider static objects of the entire environments and treat empty space as a specially-designated "object" category. Therefore, the child voxels compose a semantic occupancy map that encodes both the 3D geometry and semantic meaning of the parent voxel.

We further vectorize the semantic occupancy map and use the resulting vector as a feature descriptor of the parent voxel. The 3D environment representation e can then be represented as a 4D tensor, with dimension $X \times Y \times Z \times (M^3)$. Note that higher resolution M can better approximate the 3D shape of the environment. Our proposed HVR is thus a compact and flexible environment representation that jointly considers the 3D action location candidates, geometric and semantic information of the 3D environment.

Joint Learning of Action Category and Action Location. We present an overview of our model in Fig. 2(b). Specifically, we adopt a two-pathway network architecture. The video pathway extracts video features with an I3D backbone network $\phi(x)$, while the environment pathway extracts the global 3D environment features with a 3D convolutional network $\psi(e)$. Visual and environmental features are jointly considered for predicting the 3D action location r. We then adopt stochastic units to generate sampled action $\tilde{r}$ for selecting the local environment features relevant to the actions. Local environment features and video features are further fused together for activity recognition.

Our key idea is to utilize the 3D environment representation e for jointly modeling the action label y and 3D action location r of video clip x. We consider the action location r as a probabilistic variable, and model the action label y given input video x and environment representation e using a latent variable model. Therefore, the conditional probability $p(y|x, e)$ is given by:

$$p(y|x, e) = \int_r p(y|r, x, e)p(r|x, e)dr. \tag{1}$$

Notably, our proposed joint model has two key components. First, $p(r|x, e)$ models the 3D action location r from video input x and the 3D environment representation e. Second, $p(y|r, x, e)$ utilizes r to select a region of interest (ROI) from the environment representation e, and combines selected environment features with the video features from x for action classification. During training, our model receives the ground truth 3D action location and action label as supervisory signals. At inference time, our model jointly predicts both the 3D action location r and action label y. We now provide additional technical details in modeling $p(r|x, e)$ and $p(y|r, x, e)$.

3.2   3D Action Localization

We first introduce our 3D action localization module, defined by the conditional probability $p(r|x, e)$. Given the video pathway features $\phi(x)$ and the environment pathway features $\psi(e)$, we learn a mapping function to predict location r, which is defined on a 3D grid of candidate action locations. Note that the 3D grid is defined globally over the 3D environment scan. The mapping function

is composed of 3D convolution operations with parameters w_r and a softmax function. Thus, $p(r|x,e)$ is given by:

$$p(r|x,e) = softmax(w_r^T(\phi(x) \oplus \psi(e))), \tag{2}$$

where $\oplus$ denotes concatenation along the channel dimension. Therefore, the resulting action location r is a proper probabilistic distribution normalized in 3D space, and $r(w,d,h)$ can be considered as the expectation of video clip x happening in the spatial location (w,d,h) of the 3D environment.

In practice, we do not have access to the precise ground truth 3D action location and must rely on camera registration results as a proxy. Using a categorical distribution for $p(r|x,e)$ thus models the ambiguity of 2D to 3D registration. We follow [30, 32] to adopt stochastic units in our model. Specifically, we follow the Gumbel-Softmax and reparameterization trick from [19, 35] to adopt the following differentiable sampling mechanism:

$$\tilde{r}_{w,d,h} \sim \frac{\exp((\log r_{w,d,h} + G_{w,d,h})/\theta)}{\sum_{w,d,h} \exp((\log r_{w,d,h} + G_{w,d,h})/\theta)}, \tag{3}$$

where G is a Gumbel Distribution for sampling from a discrete distribution. This Gumbel-Softmax trick produces a "soft" sample that allows the gradients propagation to video pathway network ϕ and environment pathway network ψ. θ is the temperature parameter that controls the shape of the soft sample distribution. We set $\theta = 2$ for our model. Notably, the expectation of sampled 3D action location $E[\tilde{r}]$ can be modeled by the distribution $p(r|x,e)$ using Eq. 2.

3.3 Action Recognition with Environment Prior

Our model further models $p(y|r,x,e)$ with a mapping function $f(\tilde{r},x,e)$ that jointly considers action location r, video input x and 3D environment representation e for action recognition. Formally, the conditional probability $p(y|r,x,e)$ can be modeled as:

$$p(y|r,x,e) = f(\tilde{r},x,e) = softmax(w_p^T \Sigma(\phi(x) \oplus (\tilde{r} \otimes \psi(e)))), \tag{4}$$

where $\oplus$ denotes concatenation along channel dimension, and $\otimes$ denotes the element-wise multiplication. Specifically, our method uses the sampled action location $\tilde{r}$ for selectively aggregating environment features $\psi(e)$ and combines the aggregated environment features with video features $\phi(x)$ for action recognition. Σ denotes the average pooling operation that maps 3D feature to 2D feature, and w_p denotes the parameters of the linear classifier that maps feature vector to action prediction logits.

3.4 Training and Inference

We now present our training and inference schema. At training time, we assume a prior distribution of action location $q(r|x,e)$ is given as a supervisory signal.

$q(r|x, e)$ is obtained by registering the egocentric camera into the 3D environment (see more details in Sec.4). Note that we factorize $p(r|x, e)$ as latent variables, and based on the Evidence Lower Bound (ELBO), the resulting deep latent variable model has the following loss function:

$$\mathcal{L} = -\sum_r \log p(y|r, x, e) + KL[p(r|x, e)||q(r|x, e)], \tag{5}$$

where the first term is the cross entropy loss for action classification and the second term is the KL-Divergence that matches the predicted 3D action location distribution $p(r|x, e)$ to the prior distribution $q(r|x, e)$. During training, a single 3D action location sample $\tilde{r}$ for each input within the mini-batch will be drawn.

Theoretically, our model should sample $\tilde{r}$ from the same input multiple times and take average of the predictions at inference time. To avoid such dense sampling for high dimensional video input, we choose to directly plug in the deterministic action location r in Eq. 4. Note that the recognition function f is composed of a linear mapping function and a softmax function, and therefore is convex. Further, $\tilde{r}$ is sampled from the probabilistic distribution of 3D action location r, similar to the formulation in [29, 32], r is thus the expectation of $\tilde{r}$. By Jensen's Inequality, we have:

$$E[f(\tilde{r}, x, e)] \geq f(E[\tilde{r}], x, e) = f(r, x, e). \tag{6}$$

That being said, $f(r, x, e)$ provides an empirical lower bound of $E[f(\tilde{r}, x, e)]$, and therefore provides a valid approximation of dense sampling.

3.5   Network Architecture

For the video pathway, we adopt the I3D-Res50 network [2, 59] pre-trained on Kinetics as the backbone. For the environment pathway, we make use of a lightweight network (denoted as EnvNet), which has four 3D convolutional operations. The video features from the 3rd convolutional block of I3D-Res50 and the environment features after the 2nd 3D convolutional operation in EnvNet are concatenated for 3D action location prediction. We then use 3D max pooling operations to match the size of action location map to the size of the feature map of the 4th convolution of EnvNet for the weighted pooling in Eq.2. More implementation details can be found in our supplement.

4   Experiments and Results

4.1   Dataset and Benchmarks

Datasets. Note that existing egocentric video datasets (EGTEA [30], and EPIC-Kitchens [4] etc.) did not explicitly capture the 3D environment. We follow [46] to run ORB-SLAM on EGTEA and EPIC-Kitchens. However, less than 30% of frames can be registered, and the quality of the reconstructed point cloud is

unsatisfactory. Our empirical finding is that existing visual SfM methods can not address the naturalistic egocentric videos. In contrast, the newly-developed Ego4D [13] dataset has a subset that includes egocentric videos, high-quality 3D environment reconstructions, and 3D static objects annotation.

The subset captures 34 different indoor activities from 3 real-world living rooms, resulting in 6868 action clips. Similar to [4], we consider both *seen* and *unseen* environment splits. In the seen environment split, each environment is seen in both training and testing sets (5163 instances for training, and 1705 instances for testing). In the unseen split, all sequences from the same environment are either in training or testing (4392 instances for training, and 2476 instances for testing). As discussed in [13], the photo-realistic 3D reconstruction of the environment is obtained from the state-of-the-art dense reconstruction system [56]. Furthermore, the static 3D object meshes are annotated by painting an semantic label over the mesh polygons. The annotation includes 35 object categories plus a background class label. It is worthy noting that the static object annotations can be automated with the state-of-the-art 3D object detection algorithms.

Prior Distribution of 3D Activity Location. To obtain the ground truth of the activity location for each trimmed activity video clip, we first register the egocentric camera in the 3D environment using a RANSAC based feature matching method. Specifically, we first build a base map from the monochrome camera streams for 3D environment reconstruction using Structure from Motion [9, 50]. The pre-built base map is a dense point cloud associated with 3D feature points. We then estimate the camera pose of the video frame using active search [48]. Note that registering the 2D egocentric video frames in a 3D environment is fundamentally challenging, due to the drastic head rotation, featureless surfaces, and changing illumination. Therefore, we only consider the key frame camera registration, where enough inliers were matched with RANSAC. As introduced in Sec.3, the action location is defined as a probabilistic distribution in 3D space. Thus, we map the key frame camera location into the index of the 3D action location tensor, with its value representing the likelihood of the given action happening in the corresponding parent voxel. To account for the uncertainty of 2D to 3D camera registration, we further enforce a Guassian distribution to generate the final 3D action location ground truth.

Evaluation Metrics. For all experiments, we We follow [4, 30] to evaluate the performance of both action recognition using both Mean Class Accuracy and Top-1 Accuracy. As for 3D Action Localizatio, we consider 3D action localization as binary classification over the regular 3D grids. Therefore, we report the Precision, Recall, and F1 score on a downsampled 3D heatmap ($\times 4$ in X, Y direction, and $\times 2$ in Z direction) as in [29].

4.2   Action Understanding in Seen Environments

Our method is the first to utilize the 3D environment information for egocentric action recognition and 3D localization. Previous works have considered various environment contexts for other tasks, including 3D object detection, affordance prediction etc. Therefore, we adapt previous proposed contextual cues into our

Table 1. Comparison with other forms of environment context. Our Hierarchical Volumetric Representation (HVR) outperforms other methods by a significant margin on both action recognition and 3D action localization. The best results are highlighted with **boldface**, and the second-best results are underlined.

Method	Action Recognition		3D Action Localization		
	Mean Cls Acc	Top-1 Acc	Prec	Recall	F1
I3D-Res50	37.48	55.15	8.14	38.73	13.45
I3D+Obj	37.66	55.11	10.04	35.08	15.61
I3D+2DGround	38.69	55.37	10.88	36.19	16.73
I3D+SemVoxel	39.23	56.07	11.26	**38.77**	17.45
I3D+Affordance	39.95	55.82	11.55	35.35	17.41
Ours(HVR)	**41.64**	**56.94**	**16.71**	35.55	**22.73**

proposed joint model and design the following strong baselines (see supplementary material for the details of baseline models):

• **I3D-Res50** refers to the backbone network from [59]. We also use the network feature from I3D-Res50 for 3D action localization by adopting the KL loss.

• **I3D+Obj** uses object detection results from a pre-trained object detector [62] as contextual cues as in [10]. This representation is essentially an object-centric feature that describes the attended environment (*i.e.* where the camera wearer is facing towards), therefore 3D action location can not used for selecting surrounding environment features.

• **I3D+2DGround** projects the object information from the 3D environment to 2D ground plane. A similar representation is also considered in [46]. Note that the predicted 3D action location will also be projected to 2D ground plane to select local environment features.

• **I3D+SemVoxel** is inspired by [68], where we use the semantic probabilistic distribution of all the vertices within each voxel as a feature descriptor. Therefore, the resulting environment representation is a 4D tensor with dimension $X \times Y \times Z \times C$, where X, Y, Z represent the spatial dimension, and C denotes the number of object labels from the 3D environment mesh annotation introduced in Sec.4.

• **I3D+Affordance** follows [38] to use the afforded action distribution as feature descriptor for each voxel. The resulting representation is a 4D tensor with dimension $X \times Y \times Z \times N$, where N denotes the number of action classes.The afforded action distribution is derived from the training set.

Results. Our results on the seen environment split is listed in Table 1. Our method outperforms I3D-Res50 baseline by a large margin (4.2%/1.8% on Mean Cls Acc/Top1 Acc) on action recognition. We attribute this significant performance gain to explicitly modeling the 3D environment context. As for 3D action localization, our method outperforms I3D-50 by 9.3% – a relative improvement of **69%**. Notably, predicting the 3D action location based on video sequence alone is erroneous. Our method, on the other hand, explicitly models the 3D environment factor and thus improves the performance of 3D action localization.

Table 2. Ablation study for the 3D representation. We present the results of our method that adopts different semantic occupancy map resolution M.

Method	Action Recognition		3D Action Localization		
	Mean Cls Acc	Top-1 Acc	Prec	Recall	F1
I3D-Res50	37.48	55.15	8.14	38.73	13.45
I3D+SemVoxel	39.23	56.07	11.26	38.77	17.45
Ours ($M = 2$)	39.04	56.26	12.19	36.82	18.32
Ours ($M = 4$)	**41.64**	**56.94**	**16.71**	35.55	22.73
Ours ($M = 8$)	40.06	56.04	16.13	**39.84**	**22.96**

In subsequent sections, we will show that the performance improvement does not simply come from additional input modalities of 3D environment, but attributes to a careful design of 3D representation and probabilistic joint modeling.

Comparison on environment representation. We now compare HVR with other forms of environment representation. As shown in Table 1, I3D+Obj has minor improvement on the over all performance, while I3D+2DGround, I3D+SemVoxel and I3D+Affordance can improve the performance of action recognition and 3D localization by a notable margin. Those results suggest that the environment context (even in 2D space) plays an important role in egocentric action understanding. More importantly, our method outperforms all previous methods by at least 1.7% for action recognition and 5.3% for 3D action localization. *These results suggest that our proposed HVR is superior to a 2D ground plane representation*, and demonstrates that using the semantic occupancy map as the environment descriptor can better facilitate egocentric understanding.

4.3   Ablation Studies

We now present detailed ablation studies of our method on seen split. To begin with, we analyze the role of semantic and geometric information in our hierarchical volumetric representation (HVR). We then present an experiment to verify whether fine-grained environment context is necessary for egocentric action understanding. Furthermore, we show the benefits of probabilistic joint modeling of action and 3D action location.

Semantic Meaning and 3D Geometry. The semantic occupancy map carries both geometric and semantic information of the local environment. To show how each component contributes to the performance boost, we compare Ours with I3D+SemVoxel, where only semantic meaning is considered, in Table 2. Ours outperforms I3D+SemVoxel by a notable margin for action recognition and a large margin for 3D localization. These results suggest that semantic occupancy map is more expressive than only semantic information for action understanding, yet it has smaller impact on action recognition than 3D action localization.

Granularity of 3D Information. We further show what level of 3D environment granularity is needed for egocentric action understanding. By the definition of occupancy map, increasing the resolution M of children voxels will approximate the actual 3D shape of the environment. Therefore, we report results of

Table 3. Ablation study for joint modeling of action category and 3D action location. Our proposed probabilistic joint modeling can consistently benefit the performance on action recognition and 3D action localization

Method	Action Recognition		3D Action Localization		
	Mean Cls Acc	Top-1 Acc	Prec	Recall	F1
I3D-Res50	37.48	55.15	8.14	**38.73**	13.45
I3D+GlobalEnv	35.99	54.93	8.82	36.40	14.20
I3D+DetEnv	39.37	55.88	14.11	32.66	19.71
Ours	**41.64**	**56.94**	**16.71**	35.55	**22.73**

our method with different occupancy map resolution in Table 2. Not surprisingly, low occupancy map resolution lags behind Ours for action recognition by 2.6% ↓ and 3D action localization by 4.4% ↓, which again shows the necessity of incorporating the 3D geometric cues. Another interesting observation is that higher resolution can slightly increase the 3D action localization accuracy by 0.2%, yet decreases the performance on action recognition by 1.6% ↓. These results suggest that fine-grained 3D shape of the environment is not necessary for action recognition. In fact, higher resolution will dramatically increase the feature dimension of the environment representation, and thereby incurs more barriers to the network.

Joint Learning of Action and 3D Location. We denote a baseline model that directly fuses global environment features, extracted by the same 3D convolutional network from our method, with video features for activity reasoning as I3D+GlobalEnv. The results are presented in Table 3. I3D+GlobalEnv decreases the performance of I3D-Res50 backbone network by 1.5% ↓ /0.2% ↓ for action recognition and has marginal improvement for 3D action localization (+0.8%). We speculate that this is because only 3 types of scene reconstruction available for training may lead to overfitting. In contrast, our method makes use of the learned 3D action location to select interesting environment features associated with the action. As the action location varies among different input videos, our method can utilize the 3D environment context without running into the pitfall of overfitting, and therefore outperforms I3D+GlobalEnv by 5.7%/2.0% for action recognition and 8.5% for 3D action localization.

Probabilistic Modeling of 3D Action Location. As introduced in Sec.4, considerable uncertainty lies in the prior distribution of 3D action location, due to the challenging artifact of 2D to 3D camera registration. To verify that the probabilistic modeling can account for the uncertainty of 3D action location ground truth, we compare our method with a deterministic version of our model, denoted as DetEnv. DetEnv adopts the same inputs and network architecture as our method, except for the differentiable sampling with Gumbel-Softmax Trick. As shown in Table 3, Ours outperforms DetEnv by 2.3% for action recognition and 3.0% for 3D action localization. These results demonstrate the benefits of the stochastic units adopted in our method.

Table 4. Experimental results on unseen environment split. Our model show the capacity of better generalizing to an unseen environment with known 3D map. The best results are highlighted with **boldface**, and the second-best results are <u>underlined</u>.

Method	Action Recognition		3D Action Localization		
	Mean Cls Acc	Top-1 Acc	Prec	Recall	F1
I3D-Res50	29.24	52.22	6.20	45.14	10.90
I3D+2DObject	29.91	53.05	6.31	42.22	10.98
I3D+2DGround	30.06	<u>53.87</u>	6.95	41.27	11.90
I3D+SemVoxel	<u>30.19</u>	53.37	<u>7.03</u>	<u>43.55</u>	<u>12.11</u>
Ours	**31.55**	**55.33**	**7.50**	**44.97**	**12.86**

Remarks. To summarize, our key finding is that both 3D geometric and semantic contexts convey important information for action recognition and 3D localization. Another important take home is that egocentric understanding only requires a sparse encoding of geometric information. Moreover, without a careful model design, the 3D environment representation has minor improvement on (or even decreases) the overall performance as reported in Table 3.

4.4 Generalization to Novel Environment

We further present experiment results on the unseen environment split in Table 4. Our model outperforms all baselines by a notable margin on both action recognition and 3D action localization. Note that the affordance map requires the observation of action instances on the 3D spatial location and thus cannot be applied on the unseen split. These results suggest that explicitly modeling the 3D environment context can improve the generalization ability to unseen environments with known 3D maps. However, the performance gap is smaller in comparison to the performance boost on seen split. We speculate that this is because we only have two different types of environments for training and therefore the risk of overfitting on unseen split is further exemplified.

4.5 Discussion

Visualization of Action Location. We visualize our results on seen environment split. Specifically, we project the 3D saliency map of action location on the top-down view of the 3D environments. As shown in Fig. 3, our model can effectively localize the coarse action location and thereby select the region of interest from the global environment features for action recognition. By examining the failure cases, we found that the model may run into the failure modes when the video features are not sufficiently discriminative (*i.e.* when the camera wearer is standing close to a white wall.)
Limitation and Future Work. One limitation of our method is the requirement of high-quality 3D reconstruction with object annotations. However, we conjecture that 3D object detection algorithms [68], semantic structure from
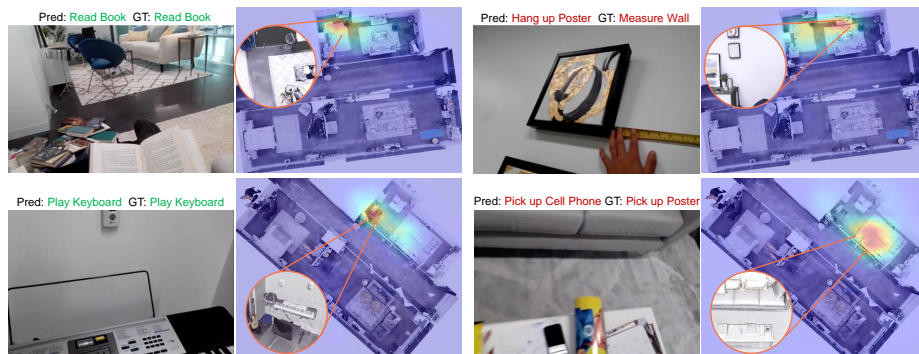
Fig. 3. Visualization of predicted 3D action location (projected on top-down view of the reconstructed 3D scene) and action labels (captions above the video frames). We present both successful and failure examples. We also show the "zoom-in" spatial region of the action location to help readers to better interpret our action localization results.

motion [26] and 3D scene graphs [1] can be used to replace the human annotation, since our current volumetric representation only adopts a low resolution semantic occupancy map as environment descriptor. We plan to explore this direction as our future work. Another limitation is the potential error in 2D to 3D camera registration, as discussed in Sec.4. Currently, only camera poses from key video frames can be robustly estimated. Our method thus does not model the location shift within the same action. We argue that camera registration can be drastically improved with the help of additional sensors (*e.g.* IMU or depth camera). Incorporating those sensors into egocentric capturing setting is an exciting future direction. In addition, our method did not consider the camera orientation. We will leave this for future efforts.

5   Conclusion

We introduced a deep model that makes use of egocentric videos and a 3D map to address the novel task of joint action recognition and 3D localization. Our key insight is that the 3D geometric and semantic context of the surrounding environment provides critical information that complements video features for action understanding. The key innovation of our model is to characterize the 3D action location as a latent variable, which is used to select the surrounding local environment features for action recognition. Our model demonstrated impressive results on seen and unseen environments when evaluated on the newly released Ego4D dataset [13]. We believe our work provides a critical first step towards understanding actions in the context of a 3D environment, and points to exciting future directions in connecting egocentric vision and 3D scene understanding for AR and HRI.

References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: ICCV (2019)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
3. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: CVPR (2017)
4. Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Computer Architecture Letters (01), 1–1 (2020)
5. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 284–298. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
6. Engelcke, M., Rao, D., Wang, D.Z., Tong, C.H., Posner, I.: Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In: ICRA (2017)
7. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: CVPR (2018)
8. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. IJCV (2014)
9. Frahm, J.M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building rome on a cloudless day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. pp. 368–381. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
10. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In: ICCV (2019)
11. Gordon, D., Kadian, A., Parikh, D., Hoffman, J., Batra, D.: Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In: ICCV (2019)
12. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR (2011)
13. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021)
14. Guan, J., Yuan, Y., Kitani, K.M., Rhinehart, N.: Generative hybrid representations for activity forecasting with no-regret learning. In: CVPR (2020)
15. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR (2011)
16. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019)
17. Henriques, J.F., Vedaldi, A.: Mapnet: An allocentric spatial memory for mapping environments. In: CVPR (2018)
18. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 789–804. Springer International Publishing, Cham (2018)
19. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)

20. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: CVPR (2013)
21. Jiang, Y., Lim, M., Saxena, A.: Learning object arrangements in 3d scenes using human context. In: ICML (2012)
22. Karthika, S., Praveena, P., GokilaMani, M.: Hololens. International Journal of Computer Science and Mobile Computing **6**(2), 41–50 (2017)
23. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
24. Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: CVPR (2019)
25. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research **32**(8), 951–970 (2013)
26. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 703–718. Springer International Publishing, Cham (2014)
27. Li, J., Wang, X., Tang, S., Shi, H., Wu, F., Zhuang, Y., Wang, W.Y.: Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In: CVPR (2020)
28. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV (2013)
29. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 639–655. Springer International Publishing, Cham (2018)
30. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. TPAMI (2021)
31. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: CVPR (2015)
32. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 704–721. Springer International Publishing, Cham (2020)
33. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4d human body capture from egocentric video via 3d scene grounding. 3DV (2021)
34. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016)
35. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: ICLR (2017)
36. Moltisanti, D., Wray, M., Mayol-Cuevas, W., Damen, D.: Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In: ICCV (2017)
37. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV (2019)
38. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: CVPR (2020)
39. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: CVPR (2020)
40. Park, H., Jain, E., Sheikh, Y.: 3d social saliency from head-mounted cameras. NeurIPS (2012)

41. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012)
42. Poleg, Y., Arora, C., Peleg, S.: Head motion signatures from egocentric videos. In: ACCV (2014)
43. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: CVPR (2014)
44. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017)
45. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
46. Rhinehart, N., Kitani, K.M.: Learning action maps of large environments via first-person vision. In: CVPR (2016)
47. Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: ICCV (2017)
48. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 752–765. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
49. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Scenegrok: Inferring action maps in 3d environments. TOG (2014)
50. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
51. Serra, G., Camurri, M., Baraldi, L., Benedetti, M., Cucchiara, R.: Hand segmentation for gesture recognition in ego-vision. In: Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices. pp. 31–36 (2013)
52. Shen, Y., Ni, B., Li, Z., Zhuang, N.: Egocentric activity prediction via event modulated attention. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 202–217. Springer International Publishing, Cham (2018)
53. Song, S., Xiao, J.: Sliding shapes for 3d object detection in depth images. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 634–651. Springer International Publishing, Cham (2014)
54. Song, S., Xiao, J.: Deep sliding shapes for amodal 3d object detection in rgb-d images. In: CVPR (2016)
55. Soo Park, H., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: CVPR (2016)
56. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
57. Sulaiman, M.Z., Aziz, M.N.A., Bakar, M.H.A., Halili, N.A., Azuddin, M.A.: Matterport: virtual tour as a new marketing approach in real estate business during pandemic covid-19. In: Proceedings of the International Conference of Innovation in Media and Visual Design (IMDES 2020). Atlantis Press. pp. 221–226 (2020)
58. Wang, D.Z., Posner, I.: Voting for voting in online point cloud object detection. In: TSS (2015)
59. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)

60. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: CVPR (2019)
61. Wu, W., Wang, Z.Y., Li, Z., Liu, W., Fuxin, L.: Pointpwc-net: Cost volume on point clouds for (self-)supervised scene flow estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 88–107. Springer International Publishing, Cham (2020)
62. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
63. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: CVPR (2018)
64. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: CVPR (2017)
65. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: Proximity learning of articulation and contact in 3D environments. In: 3DV (2020)
66. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: CVPR (2020)
67. Zhou, Y., Ni, B., Hong, R., Yang, X., Tian, Q.: Cascaded interactional targeting network for egocentric video analysis. In: CVPR (2016)
68. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: CVPR (2018)