Supplementary Materials for: 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

This is the supplementary material for our ECCV 2022 paper, titled "Egocentric Activity Recognition and Localization on a 3D Map". The contents are organized as follows.

- A Implementation Details.
- B Dataset Details.
- C Analysis of Table 1 in Main Paper
- D Additional Qualitative Results.
- E Code and Licenses.

A Implementation Details

Data Processing. We resize all video frames to the short edge size of 256. For the coarse 3D map, we adopt a resolution of $28 \times 28 \times 8$ for parent voxel, and $M = 4$ for children voxels. For training, our model takes an input of 8 frames (temporal sampling rate of 8) with a resolution of 224×224. For inference, our model samples 30 clips from a video (3 along spatial dimension and 10 in time). Each clip has 8 frames with a resolution of 224×224. We average the scores of all sampled clips for video level prediction.

Training Details. Our model is trained using SGD with momentum 0.9 and batch size 64 on 4 GPUs. The initial learning rate is 0.0375 with cosine decay. We set weight decay to 1e-4 and enable batch norm [5]. To avoid overfitting, we adopt several data augmentation techniques, including random flipping, rotation, cropping and color jittering. Our model is implemented in PyTorch. Our scource code is included with this supplement and will be made publicly available.

B Dataset Details

This section introduces details of our new egocentric video dataset. We present additional qualitative results of the key frame camera registration, and show

[⋆] This work was primarily done during an internship at Meta Reality Labs.

additional rendered images of the 3D reconstructions from our dataset. Finally, we plot the distribution of activity categories in our dataset.

Camera Registration. We provide qualitative results of key frame egocentric camera registration in Fig. 1. Specifically, we use the estimated camera pose to render the egocentric view of the 3D environment. When the camera registration fails (*e.g.* insufficient inliers are matched by RANSAC), we assign a dummy result at the world origin, which will result in a completely wrong rendered view of the 3D environment as in the last row of Fig. 1. We also visualize the matched feature points [1] to help readers better interpret the RANSAC based matching method for camera registration introduced in Sec. 4 of the main paper. Notably, registering the egocentric camera into a 3D environment based on only images remains a major challenge. The motion blur, foreground occlusion, change of illumination, and featureless surfaces in indoor capture are the main failure causing factors of the 2D-to-3D registration. To mitigate these failure cases, we only consider camera relocalization using key frames to approximate the ground truth of 3D action locations, and do not model how location might shift over time within the same action. Note that, if the entire action clips does not have one robust key frame registration result, we adopt uniform distribution for the 3D action location prior $q(r|x,e)$ during training.

3D Environment Reconstruction. Our dataset captures the 3D reconstructions of three different living rooms using SOTA dense Reconstruction system [7]. We provide more rendered images of the 3D reconstructions in Fig. 2.

Activity Distribution. We present the distribution of egocentric activities in our dataset in Fig. 3. Similar to [2,6], our dataset has a "long tailed" distribution that characterizes naturalistic human behavior. For example, the action activity category of "Pick up Book from Floor" happens 1400 times, while the action of "Put Painting" on the tail occurs only 9 times. Mean Class Accuracy provides a metric that is not biased towards frequently occurred categories, and thus is better suited than Top-1 accuracy for activity recognition on our dataset. We highlight that our model outperforms I3D baseline by 4.2% on Mean Class Accuracy.

C Analysis of Table 1 in Main Paper

Remarks on Environment Representation in Table 1. Although our method requires the annotation of static objects, we are the first to show how the 3D scene proximity can facilitate egocentric video understanding. Moreover, baseline methods like I3D+SemVoxel and I3D+2DGround both use the same static objects annotation to describe the environment context as our method. Therefore, it is a fair experiment comparison between our approach and those methods. I3D+Affordance also requires extra inputs, in the form of afforded action distribution across the entire 3D map, which directly links each voxel in the map to the most likely action for that location, a much stronger prior than our HVR representation. Note that the affordance map requires that the observation of action instances densely cover the full 3D scene (which is why I3D+Affordance cannot be applied on the unseen split). Therefore, we argue that I3D+Affordance is less scalable than our model.

Baseline Details in Table 1. I3D+SemVoxel/+Affordance adopt the same architecture as our model, and the only difference is the number of input features. I3D+2DObj adopts the late fusion strategy as [4]. I3D+2DGround adopts 2D convolutional operations for extracting 2D environment features. The extracted features are further tiled into a 4D tensor and concatenated with video features for recognition. We will add those descriptions in camera ready.

Additional Analysis. We specifically compare the activity classes, where our model significantly outperforms I3D and I3D+2DGround in Fig. 4. Interestingly, our model performs better at classes where the video features might not be discriminative enough for recognition, *e.g.* "Pick up Book from Floor" vs. "Put Book on Shelf", or "Pick up Poster" vs. "Stamp Poster". Moreover, the contextual features from 2D ground plane provide limited information for understanding those egocentric activities. We conjecture that our method makes use of environment features surrounding the predicted 3D action location to complement video features for activity recognition.

D Additional Qualitative Results

Finally, we provide additional qualitative results. As shown in Fig. 5, we present predicted 3D action location and action labels. The figure follows the same format as Fig. 3 in the main paper. Those results suggest that our model can effectively localize the action location and thereby more accurately predict the action labels. Another interesting observation is that the model may output a "diffused" heatmap, when the foreground active objects take up the majority of the video frames (right column of Fig. 5). This is because the model receives uniform prior as supervisory signals when the camera registration fails for an action clip. In these cases, our model opts for predicting a diffused heat map of action location to prevent itself from missing important environment features. In doing so, our model might still be able to successfully predict the action labels, despite the failure of camera registration.

E Code and Licenses

Our implementation is built on top of [3], which is under the Apache License[6]. We will make our code publicly available together with the camera ready paper.

References

1. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 510–517. Ieee (2012) 2

[6] https://github.com/facebookresearch/SlowFast/blob/main/LICENSE

2. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018) 2
3. Fan, H., Li, Y., Xiong, B., Lo, W.Y., Feichtenhofer, C.: Pyslowfast. https://github.com/facebookresearch/slowfast (2020) 3
4. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In: ICCV (2019) 3
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) 1
6. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. TPAMI (2021) 2
7. 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) 2

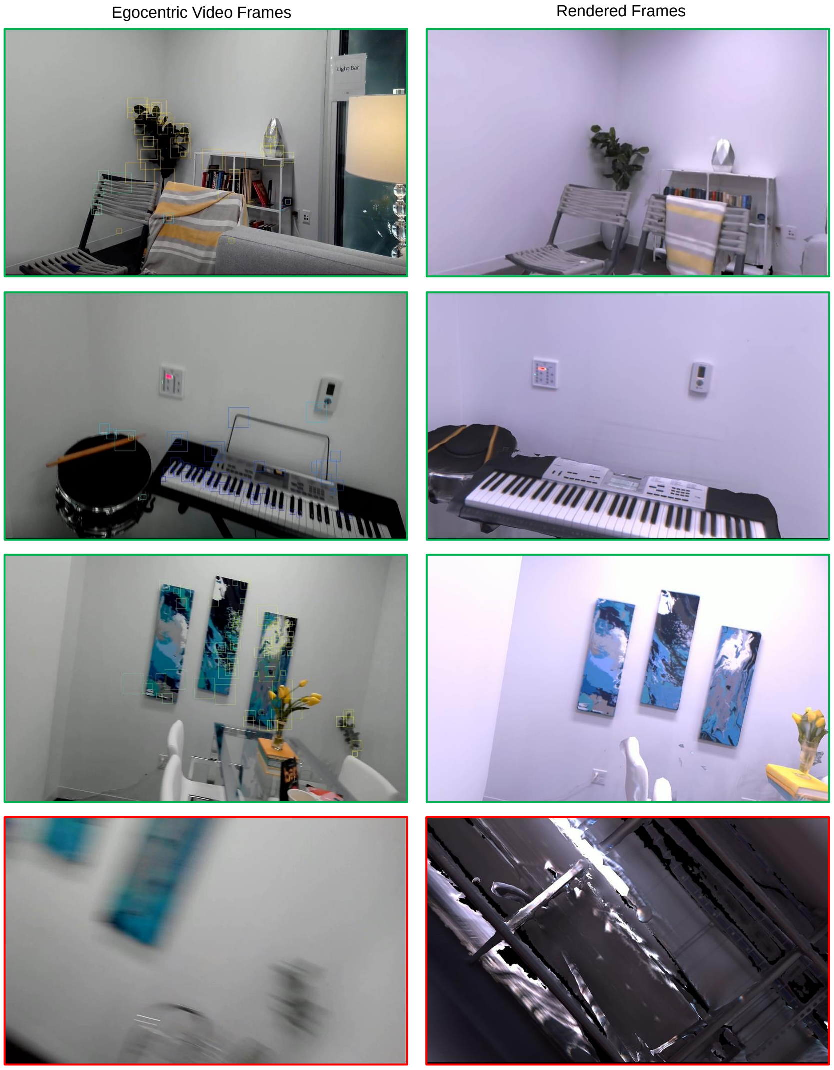

Fig. 1: Qualitative Results for Camera Registration. We present both successful cases and failure cases of the camera registration results by using the estimated camera pose to render the egocentric view of the 3D environment reconstructions.

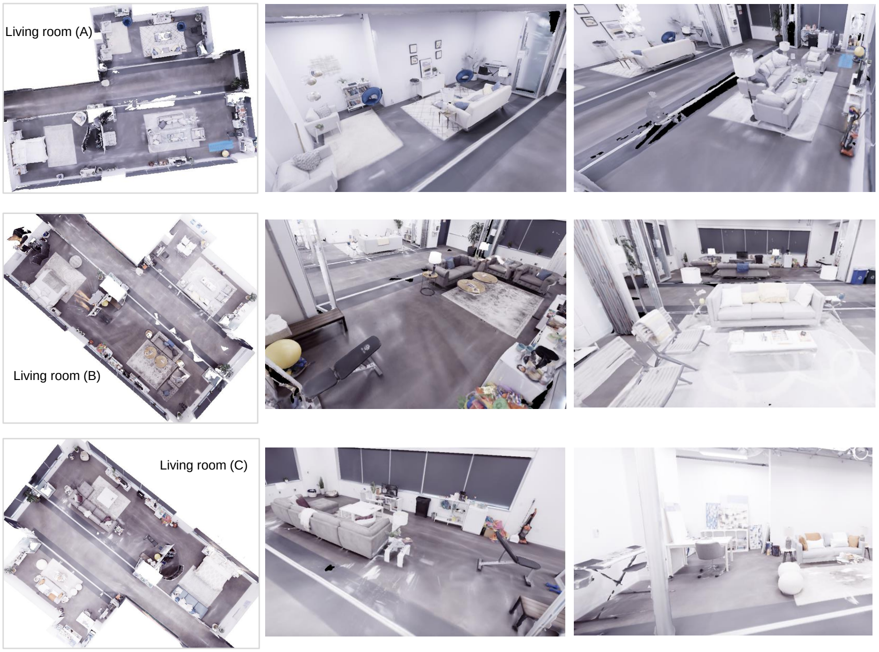

Fig. 2: Rendered views of the 3D environment reconstructions captured in our dataset.

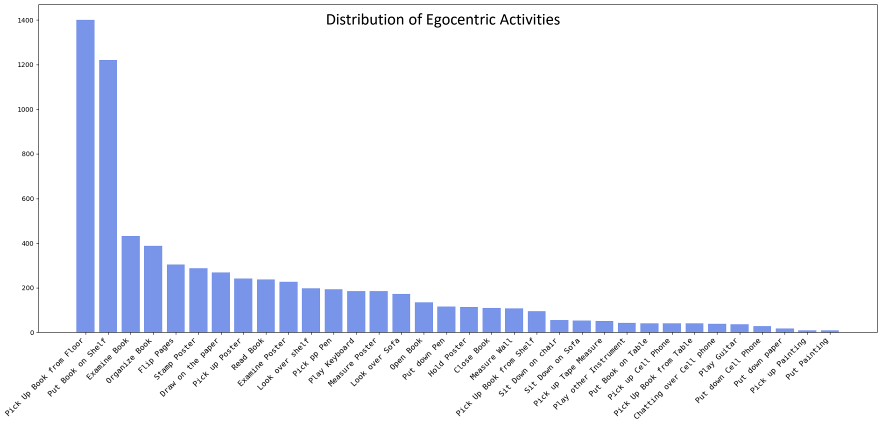

Fig. 3: Long tailed activity distribution in our dataset.

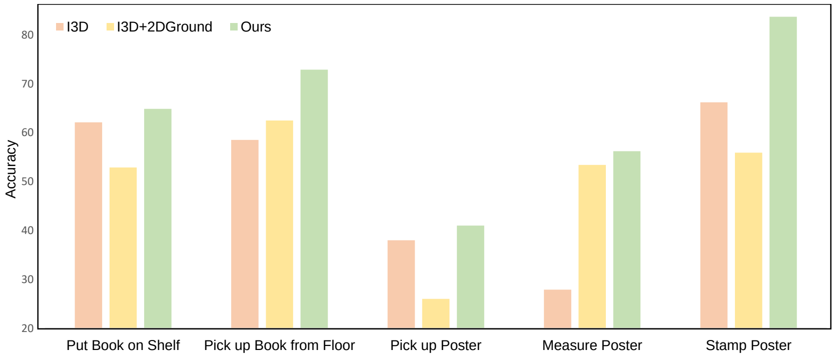

Fig. 4: A closer look at the experiment results comparison in Table 1 of main paper.

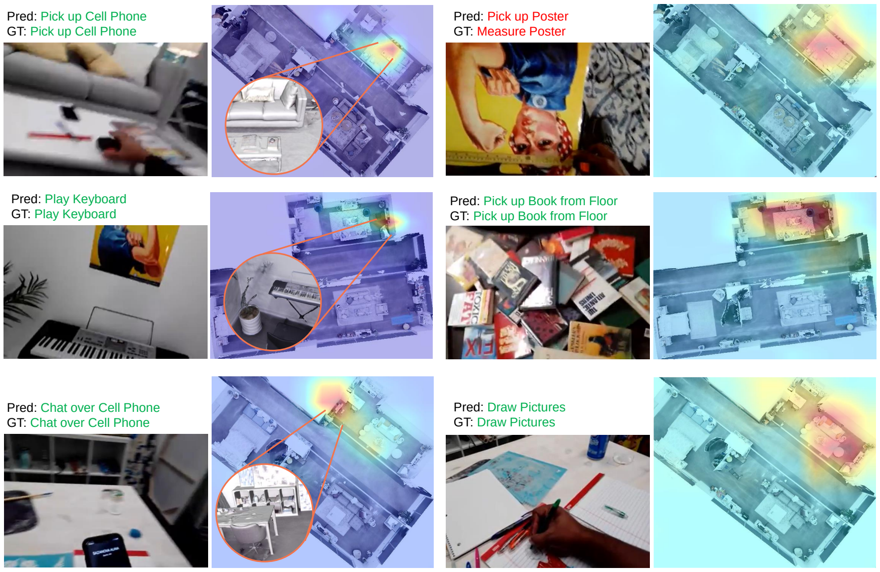

Fig. 5: Additional visualization of predicted 3D action location and action labels.

Table 1: Network architecture of our two-pathway network. We omit the residual connection in backbone ResNet-50 for simplification.

ID	Branch	Type	**Kernel Size**	**Stride**	**Output Size**	**Comments (Loss)**
			THW,(C)	THW	THWC	
1	Backbone Input Size: 8x224x224x3	Conv3D	5x7x7,64	1x2x2	8x112x112x64	
2		MaxPool1	1x3x3	1x2x2	8x56x56x64	
3		Layer1 Bottleneck 0-2	3x1x1,64 1x3x3,64 (×3) 1x1x1,256	1x1x1 1x1x1 (×3) 1x1x1	8x56x56x256	
4		MaxPool2	2x1x1	2x1x1	4x56x56x256	
5		Layer2 Bottleneck 0	3x1x1,128 1x3x3,128 1x1x1,512	1x1x1 1x2x2 1x1x1		
6		Layer2 Bottleneck 1-3	3x1x1,128 1x3x3,128 (×3) 1x1x1,512	1x1x1 1x2x2 (×3) 1x1x1	4x28x28x512	Concat with Env Feat for 3D Action Location Prediction
7		Layer3 Bottleneck 0	3x1x1,256 1x3x3,256 1x1x1,1024	1x1x1 1x2x2 1x1x1		
8		Layer3 Bottleneck 1-5	3x1x1,256 1x3x3,256 (×5) 1x1x1,1024	1x1x1 1x1x1 (×5) 1x1x1	4x14x14x1024	
9		Layer4 Bottleneck 0	3x1x1,128 1x3x3,128 1x1x1,512	1x1x1 1x2x2 1x1x1		
10		Layer4 Bottleneck 1-2	3x1x1,128 1x3x3,128 (×2) 1x1x1,512	1x1x1 1x2x2 (×2) 1x1x1	4x7x7x2048	Concat with Env Feat for Activity Recognition
11	EnvNet Input Size 8x28x28x64	Conv3d 1	3x3x3,356	2x1x1	4x28x28x256	
12		Conv3d 2	1x3x3,512	1x1x1	4x28x28x512	Concat with Video Feat for 3D Action Location Prediction
13		Action Location Branch Conv3d 1	1x3x3,512	1x2x2	4x14x14x512	
13		Action Location Branch Conv3d 2	1x3x3,1	1x1x1	4x14x14x1	KLD Loss
14		Gumbel Softmax 1 (Sampling)			4x14x14x1	Sampling 3D Action Location
15		Maxpool	2x1x1	2x1x1	4x14x14x512	
16		Conv3d 3	1x3x3,1024	1x1x1	4x14x14x1024	
17		Weighted Avg Pooling			4x14x14x1024	Guided by Sampled 3D Action Location
18		Conv3d 4	1x3x3,1024	1x3x3	4x7x7x1024	Concat with Video Feat for Activity Recognition
22	Recognition Network	Weighted Avg Pooling	4x7x7	4x7x7	1x1x1x1024	Fused Environmental and Video features
23		Fully Connected			1x1x1xN	
24		Softmax			1x1x1xN	Cross Entropy Loss