Supplementary Material for
LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding

Gen Li[1] Varun Jampani[2] Deqing Sun[2] Laura Sevilla-Lara[1]

[1]University of Edinburgh [2]Google Research

Contents

A. Framework Comparison **1**

B. Additional Experimental Details **1**
B.1. Evaluation Metrics 1
B.2. Details of the Unseen Setting 2

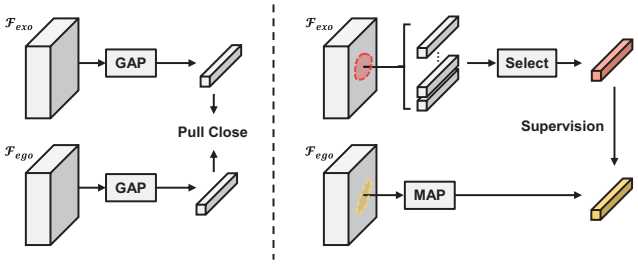
C. Additional Experimental Results **2**
C.1. Comparison on Different Scales 2
C.2. Ablation Study on Loss Function 2
C.3. Ablation Study on Hyper-parameters 3

D. Additional Visualizations **3**
D.1. Exocentric Localizations 3
D.2. Part-aware Features in DINO-ViT 4
D.3. Additional Qualitative Results 4

E. Limitations **4**

A. Framework Comparison

In Fig. 1, we show the framework comparison between state-of-the-art affordance grounding work [7, 9] and LOCATE. Previous work performs knowledge transfer by pulling close two global embeddings, and the prediction is only generated at the inference stage. Instead, we conduct part-level knowledge transfer by selecting the object part prototype from exocentric features, and utilizing it as a high-level pseudo-supervision to guide egocentric localization in an explicit manner.



(a) Previous Work (b) LOCATE

Figure 1. Comparison of LOCATE and previous work. Previous affordance grounding work [7, 9] performs knowledge transfer in a global and implicit manner. In contrast, LOCATE conducts knowledge transfer in a more localized and explicit manner. (GAP denotes global average pooling, and MAP is masked average pooling).

B. Additional Experimental Details

B.1. Evaluation Metrics

Different from the semantic segmentation task that uses the binary mask as ground truth, the GT for affordance grounding is a probability distribution that indicates the affordance area, i.e., "action possibilities". Following previous work [3, 5–7, 9], we use KLD, SIM, and NSS as the metrics to evaluate the prediction performance. KLD, SIM, and NSS are used to measure the difference, similarity, and correspondence between two probability distributions, respectively. Here, we detail the calculation of each metric. Specifically, we first feed the prediction $\mathcal{P} \in \mathbb{R}^{H \times W}$ and ground truth $\mathcal{M} \in \mathbb{R}^{H \times W}$ to a min-max normalization. To compute KLD and SIM, input maps are divided by the sum of all elements:

$$\hat{\mathcal{P}}_i = \mathcal{P}_i / \sum \mathcal{P}, \quad \hat{\mathcal{M}}_i = \mathcal{M}_i / \sum \mathcal{M}. \tag{1}$$

Then, KLD and SIM are calculated as

$$\mathrm{KLD}(\hat{\mathcal{M}} \parallel \hat{\mathcal{P}}) = \sum_i \hat{\mathcal{M}}_i \cdot log(\frac{\hat{\mathcal{M}}_i}{\hat{\mathcal{P}}_i}), \tag{2}$$

$$\mathrm{SIM}(\hat{\mathcal{P}}, \hat{\mathcal{M}}) = \sum_i min(\hat{\mathcal{P}}_i, \hat{\mathcal{M}}_i). \tag{3}$$

For NSS, the input maps are first processed as follows:

$$\bar{\mathcal{M}} = \mathbb{1}(\mathcal{M} > 0.1), \quad \bar{\mathcal{P}} = \frac{\mathcal{P} - \mu(\mathcal{P})}{\sigma(\mathcal{P})}, \tag{4}$$

Train	apple, badminton racket, baseball, baseball bat, basketball, bench, book, bottle, bowl, carrot, cell phone, chair, couch, discus, fork, frisbee, hammer, hot dog, javelin, keyboard, knife, microwave, motorcycle, orange, oven, punching bag, rugby ball, scissors, skateboard, snowboard, suitcase, surfboard, tennis racket, toothbrush, wine glass
Test	axe, banana, bed, bicycle, broccoli, camera, cup, golf clubs, laptop, refrigerator, skis, soccer ball

Table 1. Training and test object categories under the unseen setting.

	Method	Big			Middle			Small		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Seen	EIL [8]	1.047	0.461	0.389	1.794	0.284	0.710	3.057	0.123	0.231
	SPA [10]	5.745	0.317	0.222	4.990	0.228	0.440	6.076	0.118	0.297
	TS-CAM [4]	1.039	0.424	0.166	1.814	0.248	0.401	2.652	0.132	0.352
	Hotspots [9]	0.986	0.448	0.408	1.738	0.265	0.672	2.587	0.149	0.683
	Cross-view-AG [7]	0.766	0.533	0.652	1.485	0.322	1.040	2.373	0.175	0.927
	Cross-view-AG+* [6]	0.787	0.521	0.660	1.481	0.314	1.089	2.381	0.167	0.959
	LOCATE (Ours)	**0.676**	**0.580**	**0.706**	**1.178**	**0.390**	**1.316**	**2.029**	**0.216**	**1.349**
Unseen	EIL [8]	1.199	0.393	0.271	1.906	0.246	0.482	3.082	0.113	0.116
	SPA [10]	8.299	0.259	0.254	6.938	0.186	0.333	7.784	0.095	0.144
	TS-CAM [4]	1.238	0.351	0.072	1.970	0.208	0.236	2.766	0.113	0.124
	Hotspots [9]	1.015	0.425	0.548	1.872	0.242	0.605	2.693	0.134	0.544
	Cross-view-AG [7]	0.884	0.500	0.728	1.595	0.303	0.945	2.558	0.147	0.692
	Cross-view-AG+* [6]	0.867	0.485	0.776	1.658	0.279	0.988	2.630	0.133	0.754
	LOCATE (Ours)	**0.571**	**0.629**	**0.956**	**1.302**	**0.373**	**1.257**	**2.223**	**0.189**	**1.071**

Table 2. Comparison to state-of-the-arts on different object scales. The test set is divided into three subsets (Big, Middle and Small) based on the ratio of the mask to the image. The **best** and second-best results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better). The symbol * indicates that we reproduce the results using the official code.

where $\mathbb{1}(\cdot)$ is an indicator function, μ and σ denote the arithmetic mean and standard deviation of $\mathcal{P}$, respectively. NSS is then computed as the average normalized prediction at binary GT locations:

$$\mathrm{NSS}(\bar{\mathcal{P}}, \bar{\mathcal{M}}) = \frac{1}{\sum \bar{\mathcal{M}}} \sum_i \bar{\mathcal{P}} \cdot \bar{\mathcal{M}}_i. \tag{5}$$

B.2. Details of the Unseen Setting

For the unseen setting in the AGD20K dataset [7], there are 35 object classes in the training set and 12 classes in the test set. It is worth noting that there is no object category intersection between training and test sets, so the model can learn how humans interact with objects from the training set and generalize the ability to novel objects in the test set. In Table 1, we show the object categories in training and test set, respectively.

C. Additional Experimental Results

C.1. Comparison on Different Scales

To investigate the effect of different affordance region scales on the model, we follow [7] to split the test set into three subsets: Big, Middle and Small. The egocentric images in the "Big" subset have large affordance regions (the proportion of the mask to the image content is larger than 0.1), while the "Small" subset contains samples with fairly small affordance region (mask ratio is below 0.03), which is challenging to make accurate prediction. The remaining samples will be classified to the "Middle" subset. The results are shown in Table 2, LOCATE achieves the best performance among all the other methods on all scales and metrics.

C.2. Ablation Study on Loss Function

In LOCATE, we use cosine embedding loss to perform the supervision, which pulls the egocentric embedding towards the direction of the selected prototype. To explore the impact of different objective functions, we show the perfor-

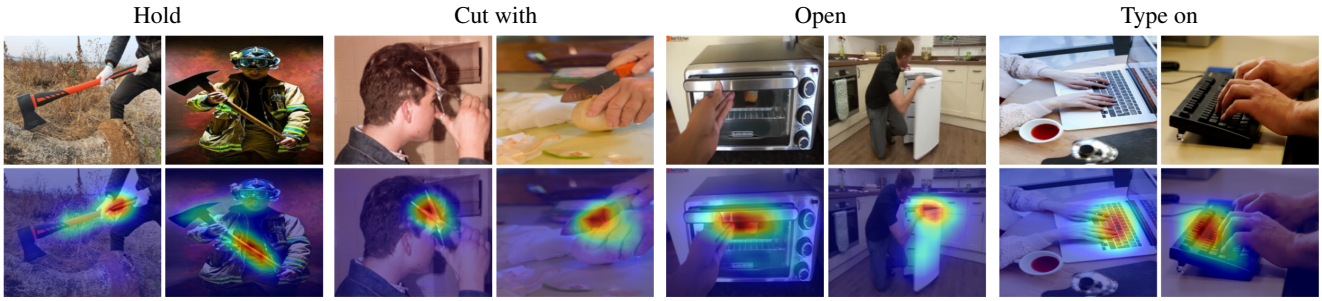


Figure 2. Localization maps for exocentric images.

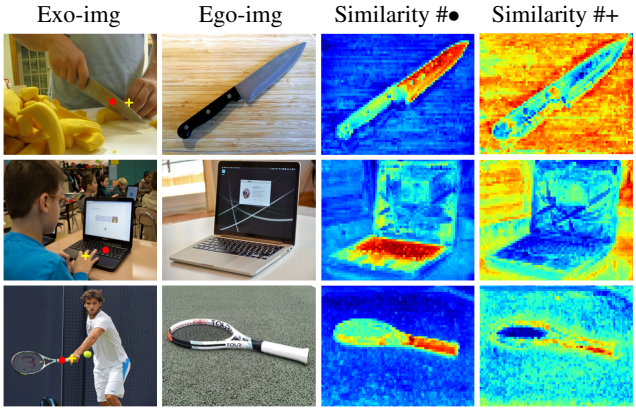


Figure 3. Similarity maps computed between exocentric embeddings corresponding to the dot/cross and all egocentric features. Here dots and crosses are placed on positions of object parts and humans, respectively.

Loss	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
MSE	1.395	0.357	1.112	1.648	0.301	1.033
Triplet	1.281	0.399	1.114	1.426	0.382	1.121
Cos†	1.240	0.400	1.156	1.418	0.370	1.148
Cos	1.226	0.401	1.177	1.405	0.372	1.157

Table 3. Ablation study on the choice of loss functions. Cos† denotes the cosine embedding loss without margin.

mance with different loss functions in Table 3. For triplet loss, we select the prototype with the lowest PartIoU as the negative example. Results show that cosine embedding loss achieves the best results, and the margin can compensate for domain gaps to further improve performance.

C.3. Ablation Study on Hyper-parameters

There are two main hyper-parameters in LOCATE. The first one is τ which controls the portion of extracted exocentric feature embeddings. A larger τ leads to more localized extraction and fewer embeddings. The other one is μ, which indicates the confidence that the selected prototype represents the object part. Ablation results of these two hyper-parameters are displayed in Table 4 and Table 5, respectively. Our model is not sensitive to the choice of hyper-parameters, as results only vary within a small range. We set (τ, μ) to 0.6 as the final setting.

τ	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0.4	1.232	0.397	1.178	1.409	0.368	1.179
0.5	1.229	0.400	1.177	1.405	0.370	1.165
0.6	1.226	0.401	1.177	1.405	0.372	1.157
0.7	1.226	0.400	1.176	1.414	0.372	1.140
0.8	1.239	0.400	1.159	1.423	0.373	1.125

Table 4. Ablation study on the localization map threshold τ.

μ	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0.4	1.228	0.399	1.180	1.406	0.369	1.161
0.5	1.232	0.399	1.177	1.404	0.371	1.161
0.6	1.226	0.401	1.177	1.405	0.372	1.157
0.7	1.235	0.400	1.163	1.405	0.372	1.167
0.8	1.230	0.400	1.163	1.411	0.372	1.149

Table 5. Ablation study on the PartIoU threshold μ.

D. Additional Visualizations

D.1. Exocentric Localizations

In LOCATE, we first extract feature embeddings from highly activated positions in the exocentric localization maps, allowing the model to focus more on object parts. In Fig. 2, we show some examples of exocentric localization maps. We find these maps mainly focus on the interaction regions, which contain strong affordance-specific knowledge of how humans interact with objects.

D.2. Part-aware Features in DINO-ViT

Our framework is built on the deep feature of a self-supervised vision transformer (DINO-ViT [2]), whose part-aware features can provide good semantic correspondences across images [1]. We provide some illustrations in Fig. 3. It can be seen that the features of DINO-ViT can help find the object parts involved in the exocentric interaction.

D.3. Additional Qualitative Results

In Fig. 4, we show more qualitative comparison with state-of-the-art methods. In the seen setting, state-of-the-art methods can locate the general affordance area, but the predicted heatmaps are very coarse with blurred boundaries. In comparison, LOCATE performs much better with more part-focused results. As for the unseen setting, most state-of-the-art approaches often locate the wrong affordance region, while our results consistently show better performance.

E. Limitations

We note that LOCATE has several limitations: (1) The performance of small objects with the affordance "holding" is generally poor, as the corresponding object parts in the exocentric images are often fully occluded, e.g., the handle of a knife. This could be potentially addressed by learning from human-object interaction videos, which can better eliminate the effects of interaction diversity and occlusion. (2) While LOCATE identifies matching object parts based on DINO-ViT features, we have observed that these features can be sensitive to factors such as texture, shadow, and lighting, which may lead to inconsistent clustering results. To improve the clustering stability, future work could introduce random crop and flip augmentations, as suggested in [1]. (3) Egocentric images in the AGD20K typically focus on a single instance and are primarily object-centric, while real-life images can be more cluttered and complex.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For*, 2022. 4

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 4

[3] Kuan Fang, Te Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2Vec: Reasoning Object Affordances from Online Videos. *CVPR*, 2018. 1

[4] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2021. 2

[5] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 1

[6] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *arXiv preprint arXiv:2208.13196*, 2022. 1, 2, 5

[7] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. *arXiv preprint arXiv:2203.09905*, 2022. 1, 2, 5

[8] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *CVPR*, 2020. 2

[9] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. *ICCV*, 2019. 1, 2, 5

[10] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *CVPR*, 2021. 2

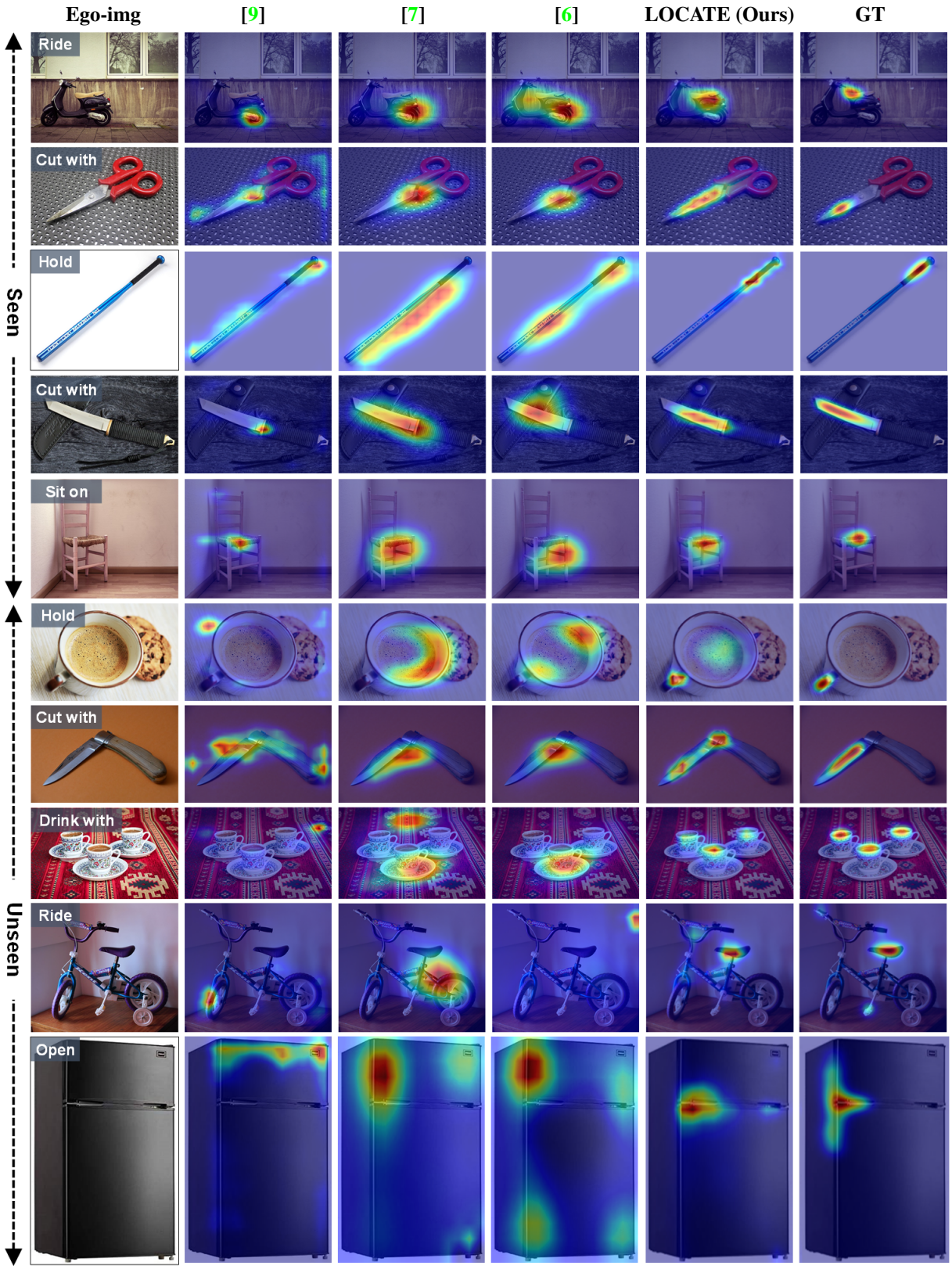


Figure 4. Qualitative comparison between LOCATE and state-of-the-art affordance grounding methods (Hotspots [9], Cross-view-AG [7], and Cross-view-AG+ [6]) in both seen and unseen settings.