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

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

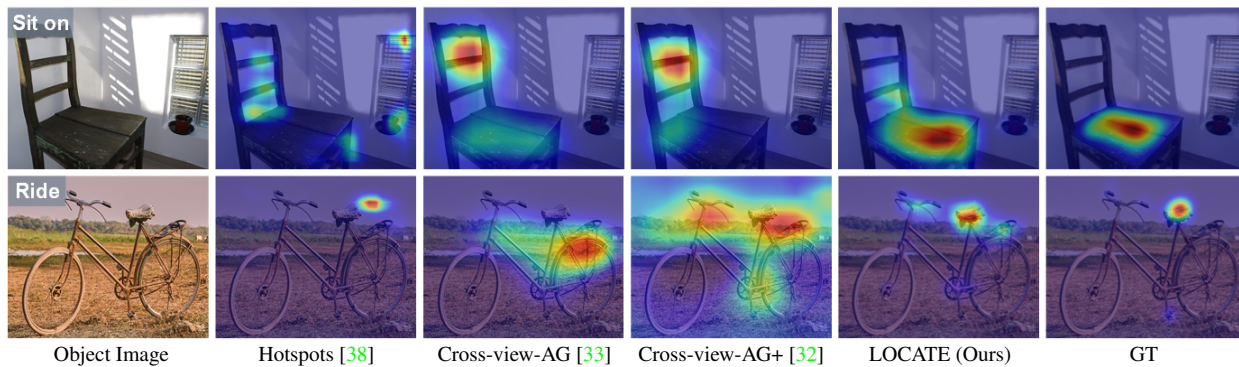[1]University of Edinburgh   [2]Google Research

Figure 1. State-of-the-art methods in weakly supervised affordance grounding often fail to make accurate predictions for objects with complex structures, e.g., chairs and bicycles. To address this, our model (LOCATE) focuses on localizing and transferring features of object parts, which is able to produce more accurate results.

Abstract

Humans excel at acquiring knowledge through observation. For example, we can learn to use new tools by watching demonstrations. This skill is fundamental for intelligent systems to interact with the world. A key step to acquire this skill is to identify what part of the object affords each action, which is called affordance grounding. In this paper, we address this problem and propose a framework called LOCATE that can identify matching object parts across images, to transfer knowledge from images where an object is being used (exocentric images used for learning), to images where the object is inactive (egocentric ones used to test). To this end, we first find interaction areas and extract their feature embeddings. Then we learn to aggregate the embeddings into compact prototypes (human, object part, and background), and select the one representing the object part. Finally, we use the selected prototype to guide affordance grounding. We do this in a weakly supervised manner, learning only from image-level affordance and object labels. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods by a large margin on both seen and unseen objects.[1]

---

[1]Project page: https://reagan1311.github.io/locate.

1. Introduction

A fundamental skill of humans is learning to interact with objects just by observing someone else performing those interactions [5]. For instance, even if we have never played tennis, we can easily learn where to hold the racket just by looking at a single or few photographs of those interactions. Such learning capabilities are essential for intelligent agents to understand what actions can be performed on a given object. Current visual systems often focus primarily on recognizing *what* objects are in the scene (passive perception), rather than on *how* to use objects to achieve certain functions (active interaction). To this end, a growing number of studies [3, 12, 26, 47] have begun to utilize affordance [17] as a medium to bridge the gap between passive perception and active interaction. In computer vision and robotics [2, 20], affordance typically refers to regions of an object that are available to perform a specific action, e.g., a knife handle affords holding, and its blade affords cutting.

In this paper, we focus on the task of affordance grounding, i.e, locating the object regions used for a given action. Previous methods [10, 12, 14, 37, 39] have often treated affordance grounding as a fully supervised semantic segmentation task, which requires costly pixel-level annotations. Instead, we follow the more realistic setting [32, 33, 38]
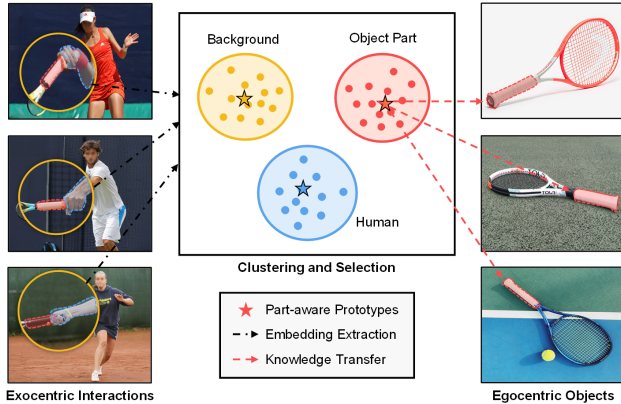
Figure 2. General illustration of LOCATE. We first extract the embeddings under the region of interest where the exocentric interactions are happening, and then split these embeddings into several clusters. In the end, the prototype of the object-part cluster is selected to supervise the egocentric affordance grounding.

where the task is learning object affordances by observing human-object interaction images. That is, given some interaction images, such as those in Fig. 2, along with the corresponding label (e.g., "hold"), the aim is to learn affordance grounding on the novel instances of that object. This is a weakly-supervised problem setting where only the image-level labels are given without any per-pixel annotations. Concretely, given several third-person human-object interaction images (exocentric) and one target object image (egocentric), our goal is to extract affordance knowledge and cues from exocentric interactions, and perform affordance grounding in the egocentric view by using only affordance labels.

There are several key challenges underlying the problem of affordance grounding. The first is due to the nature of the supervision, where only image-level affordance labels are given, being a weakly supervised problem. Here, the system needs to automatically reason about affordance regions just from classification labels. Second, human-object interactions often introduce heavy occlusion of object parts by interacting humans. In other words, the object part that the system needs to predict for a particular affordance (e.g., a mug handle for the "holding" affordance) in an exocentric image can often be the part that is occluded (e.g., by hands). Third, interactions are of great diversity. The way humans interact with objects varies across individuals resulting in diverse egocentric interaction images. Lastly, there is a clear domain gap between exocentric and egocentric images where the former have clutter, occlusion etc., and the latter are cleaner (e.g., in Fig. 2). This makes affordance knowledge transfer particularly challenging.

In this work, we propose a framework called LOCATE that addresses these core challenges by locating the exact

object parts involved in the interaction from exocentric images and transferring this knowledge to inactive egocentric images. Refer to Fig. 2 for the illustration. Specifically, we first use the class activation mapping (CAM) [51] technique to find the regions of human-object-interaction in exocentric images. Despite being trained for the interaction recognition task, we observe that CAM can generate good localization maps for interaction regions. We then segment this region of interest further into regions corresponding to human, object part, and background. We do this by extracting embeddings and performing k-means clustering to obtain several compact prototypes. Next, we automatically predict which of these prototypes corresponds to the object part relevant to the affordance. To this end, we propose a module named PartSelect that leverages part-aware features and attention maps from a self-supervised vision transformer (DINO-ViT [6]) to obtain the desired prototype. Finally, we use the object-part prototype as a high-level pseudo supervision to guide egocentric affordance grounding.

Our contributions can be summarized as follows. (1) We propose a framework called LOCATE that extracts affordance knowledge from weakly supervised exocentric human-object interactions, and transfers this knowledge to the egocentric image in a localized manner. (2) We introduce a novel module termed PartSelect to pick affordance-specific cues from human-object interactions. The extracted information is then used as explicit supervision to guide affordance grounding on egocentric images. (3) LOCATE achieves state-of-the-art results with far fewer parameters and faster inference speed than previous methods, and is able to locate accurate affordance region for unseen objects. See Fig. 1 for examples of our results and comparison to state-of-the-art.

2. Related Work

Visual Affordance Learning. Affordances are actively studied in robotics and computer vision due to their great potential to bridge the gap between perception and action. In general, affordance-related research in computer vision aims to localize objects or object parts where specific actions can be performed. Earlier work [10, 12, 27, 36] mainly focused on the fully supervised setting. Due to the scarcity of affordance datasets, this work normally required costly manual annotations to construct large-scale datasets [10, 37, 39, 42]. To alleviate the annotation and sensing requirement, new research [33, 38, 41, 42] explored acquiring affordance information in the weakly supervised setting. In particular, Sawatzky et al. [42] proposed a convolutional network to tackle affordance detection using only a few keypoint annotations, and Nagarajan et al. [38] introduced a method to infer the object affordance regions by directly learning from human-object interaction videos. In this work, we also focus on learning affordances from weak

supervision, i.e., several human-object interaction images and corresponding image-level affordance labels.

Weakly Supervised Affordance Grounding. Unlike object localization that aims to find where the object is, in affordance grounding we focus on locating functional regions of objects, which is a more fine-grained localization task. Existing weakly supervised object localization [16, 25, 34, 40, 49, 50] and semantic segmentation methods [8, 9, 28, 48, 52] are mainly based on class activation mapping (CAM) [51], the seminal work that has greatly facilitated the development of weakly supervised learning. Similarly, most affordance grounding research [31, 33, 38] also adopted CAM or its variants [7, 43] to generate activation maps as prediction. However, previous work only used CAM in the inference stage, lacking proper supervision. In contrast, to provide explicit guidance at the training stage, we utilize CAM to produce localization maps in the forward pass and supervise it in a prototypical learning fashion.

Knowledge Distillation across Domains. Knowledge distillation [23] is a process of transferring knowledge from a network to another, where inputs normally come from the same domain. Recently, many studies [14, 29, 33, 38, 44] have begun to pay more attention to knowledge distillation in different domains. This work transferred representations from the domain with numerous datasets to the other one that typically has low dataset scale and diversity, so that the model can benefit from the large quantity of data in the first domain. For instance, Li et al. [29] proposed a framework to learn egocentric video representations from large-scale third-person video datasets, and Luo et al [33] devised a knowledge transfer architecture to transfer affordance-specific features from exocentric view to egocentric view. In this paper, we follow the same setting as [33], using exocentric (third-person) human-object interaction images as the learning target. The objective of this work is to observe where and how humans interact with an object when doing a specific action, and then transfer the knowledge onto the target object.

Self-supervised Vision Transformer. Self-supervised Vision Transformers (ViT) [13], such as DINO [6], MAE [21], and BEiT [4], have demonstrated immense potential in unsupervised dense prediction tasks. Specifically, the features extracted from the self-attention layer exhibit the ability to separate different objects and generate reliable pseudo segmentation maps without requiring any manual annotations. Building on this property, recent work [15, 18, 19, 35, 45, 46] has achieved exceptional results in unsupervised segmentation and localization. Notably, Amir et al. [1] demonstrated that the pretrained features extracted from DINO-ViT encode fine-grained semantic information, which can yield excellent performance for part co-segmentation and semantic correspondence. Inspired by this work, we leverage DINO-

ViT features to identify matching object parts between exocentric and egocentric images.

3. Method

Given several exocentric interaction images and one egocentric object image, our goal is to extract affordance-related knowledge from exocentric interactions, and transfer it to egocentric images so that the affordance region can be located even for an inactive object. During training, the only supervision available are image-level affordance labels. In the inference stage, taking an egocentric image and an affordance label as input, the model needs to predict the corresponding affordance region.

The core idea of our approach is to exclude distracting information, e.g., human and background, when extracting affordance-specific features from the exocentric view, and perform fine-grained part-level knowledge transfer from exocentric images to egocentric ones. To this end, we set up the framework LOCATE to transfer the knowledge in three steps (See Fig. 3). First, we utilize CAM to generate localization maps for exocentric images, and extract corresponding feature embeddings with high activation in the localization maps (Sec. 3.1). Then, we propose PartSelect that leverages part-aware deep features to remove irrelevant information while preserving embeddings that can represent affordance cues (Sec. 3.2). Finally, we use the output from PartSelect to supervise the egocentric affordance grounding in an explicit manner (Sec. 3.3).

3.1. Locating Interaction Regions

To determine where neural networks focus on for recognition, we adopt the technique of CAM [51] to generate class-aware localization maps, which has been widely used in weakly supervised tasks. The vanilla CAM generates localization maps as a post-processing step that cannot be guided during training. However, our goal is to extract affordance-specific cues from exocentric images, and use these cues as explicit supervision for the egocentric view. Therefore, in order to obtain localization maps during the training phase, we produce class-specific feature maps instead by adding a class-aware convolution layer, which has proven to be identical to the generation process in CAM [50]. Specifically, for input images $\{I_{exo}, I_{ego}\}$ ($I_{exo} = \{I_1, I_2, ..., I_N\}$), we first extract deep features $\{\mathcal{F}_{exo}, \mathcal{F}_{ego}\} \in \mathbb{R}^{D \times H \times W}$ using a network ϕ. In our case, ϕ is a self-supervised vision transformer (DINO-ViT), whose features are part-aware and provide good part-level correspondences. We then generate localization maps $\mathcal{P}$ and classification scores z as follows:

$$\mathcal{P} = \psi_{cam}(\mathcal{F}) \in \mathbb{R}^{C \times H \times W}, \quad z = \mathrm{GAP}(\mathcal{P}) \in \mathbb{R}^{C}, \quad (1)$$

where ψ_{cam} starts with a projection layer consisting of a feed-forward layer followed by two convolutions to fine-
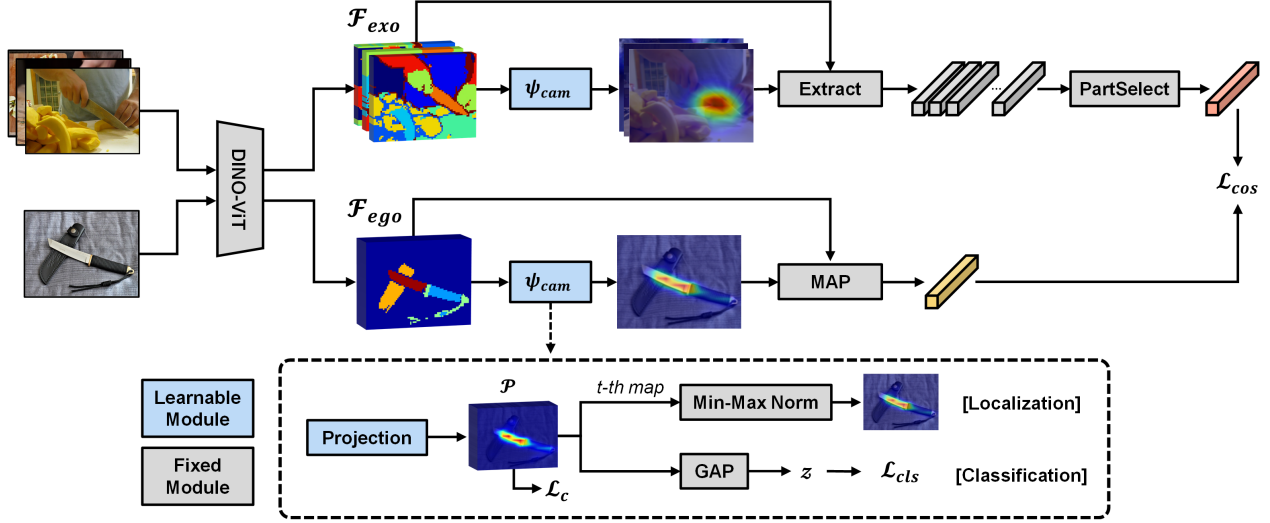
Figure 3. Overview of the proposed LOCATE framework. It achieves part-level knowledge transfer in three steps: 1) locating interaction regions with ψ_{cam} (Sec. 3.1), 2) object-part embedding selection with PartSelect (Sec. 3.2), and 3) part-level knowledge transfer with L_{cos} (Sec. 3.3). Details for PartSelect are shown in Fig. 4. At test time, only the egocentric branch is maintained.

tune features for the HOI recognition task, i.e., recognizing actions shown in the exocentric images. Then a 1×1 class-aware convolution layer is added to yield localization maps, converting the number of channels to C, where C denotes the number of total interaction categories. Therefore, each map $\mathcal{P}_c \in \mathbb{R}^{H \times W}$ represents the network activation for the c-th interaction. Next, $\mathcal{P}$ is fed to a global average pooling (GAP) layer to obtain classification scores z, which are used to calculate cross-entropy loss $\mathcal{L}_{cls}$ for optimization.

We notice that localization maps for the exocentric branch $\mathcal{P}_{exo}$ concentrate on the interaction areas, i.e, where the action takes place. Since interaction areas carry strong affordance information, we therefore aim to collect embeddings from the high activation regions in exocentric localization maps. Specifically, we first extract the localization map corresponding to ground-truth class, and conduct min-max normalization to constrain activation values to $[0, 1]$. After that, we set a threshold τ to control the number of extracted embeddings, therefore embeddings with activation value greater than τ in the localization map will be extracted from deep features. For multiple exocentric images, embeddings are extracted separately to produce $f_{exo} = [f_1, ..., f_N]$, each f_n containing a different number of embeddings. All embeddings are then concatenated together $f_{exo} \in \mathbb{R}^{L \times D}$, where L denotes the number of embeddings.

3.2. Object-Part Embedding Selection

In general, interaction areas are composed of human, object part, and background. Our objective is to eliminate the interference information, and purely deliver embeddings representing the object part to guide the egocentric branch.

In consequence, we design PartSelect to choose affordance-related embeddings from exocentric branch. PartSelect is illustrated in Fig. 4a. We first perform k-means clustering to get K compact prototypes $\boldsymbol{p} \in \mathbb{R}^{K \times D}$ from extracted exocentric embeddings. Next, we compute the cosine distance between each prototype and the egocentric deep features $\mathcal{F}_{ego}$ to get similarity maps $\mathcal{S} \in \mathbb{R}^{K \times H \times W}$:

$$S^{k,u,v} = \frac{\boldsymbol{p}_k \cdot \mathcal{F}_{ego}^{u,v}}{\|\boldsymbol{p}_k\| \, \|\mathcal{F}_{ego}^{u,v}\|}. \tag{2}$$

Owing to the fine-grained semantic information of DINO-ViT deep features, embeddings of the same object parts bear high similarity.

To distinguish which prototype stands for the object part, we aggregate the self-attention maps from the last layer of the DINO-ViT to generate a saliency map $\mathcal{A} \in \mathbb{R}^{H \times W}$ for the egocentric image. Given the saliency map and similarity maps, we introduce a metric $\gamma \in [0, 1]$ termed PartIoU to measure if a prototype carries object part information. The PartIoU for the k-th prototype is defined as follows:

$$\gamma = \frac{1}{2} \frac{\overline{\mathcal{S}_k} \cap \overline{\mathcal{A}}}{\overline{\mathcal{S}_k}} + \frac{1}{2} \frac{\overline{\mathcal{A}}}{\overline{\mathcal{S}_k} \cup \overline{\mathcal{A}}}, \tag{3}$$

where $\overline{\mathcal{S}_k}, \overline{\mathcal{A}} \in \{0, 1\}^{H \times W}$ are binary masks, we set the threshold as the average of each map to perform binarization. The motivation of PartIoU is fairly straightforward, if $\overline{\mathcal{S}}$ belongs to a portion of $\overline{\mathcal{A}}$, then the intersection of $\overline{\mathcal{S}}$ and $\overline{\mathcal{A}}$ should equal $\overline{\mathcal{S}}$ itself, while the union of the two masks should be identical to $\overline{\mathcal{A}}$. Finally, when the maximum PartIoU among K prototypes is above a threshold μ, PartSelect will output the prototype with the largest PartIoU as

(a) Illustration of PartSelect.

Exo-img   Object Part   Background   Human
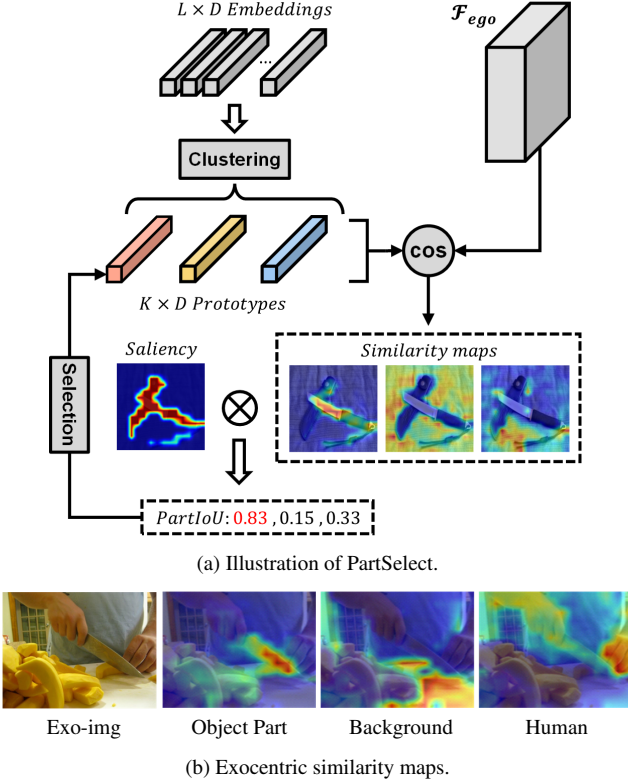
(b) Exocentric similarity maps.

Figure 4. (a) PartSelect picks the object-part prototype through clustering and selection, where $\otimes$ denotes the calculation for PartIoU score. (b) The similarity maps between prototypes and exocentric features confirm our statement that each prototype represents the object part, background, and human.

the object-part representation. Otherwise, no prototype will be selected for the next step. In Fig. 4b, we visualize the similarity maps between prototypes and exocentric features to demonstrate that the extracted embeddings are clustered into human, object part, and background.

3.3. Part-Level Knowledge Transfer

With the help of PartSelect, we find the prototype f_{op} that represents the object part. We then leverage it to perform supervision for egocentric localization maps $\mathcal{P}_{ego}$. Concretely, we first perform masked average pooling (MAP) between the normalized localization map and extracted deep features to aggregate into one embedding:

$$f_{ego} = \frac{\sum_{u=1,v=1}^{W,H} \mathcal{F}_{ego}^{u,v} \mathcal{P}_{ego}^{t,u,v}}{\sum_{i=1,j=1}^{W,H} \mathcal{P}_{ego}^{t,u,v}} \in \mathbb{R}^D, \qquad (4)$$

where t denotes the ground-truth category. Then, a cosine embedding loss is applied to pull the embedding f_{ego} to-

wards the direction of f_{op}:

$$\mathcal{L}_{cos} = \max(1 - \frac{f_{op} \cdot f_{ego}}{\|f_{op}\| \|f_{ego}\|} - \alpha, 0), \qquad (5)$$

as the two embeddings come from different domains, we thereby add α as a margin to compensate the domain gap.

In addition, since the affordance region typically denotes a portion of an object, we can thus impose a geometry loss to regulate its distribution. Inspired by the co-part segmentation work [24], we add a concentration loss to encourage egocentric localization maps to form a concentrated and connected component. The concentration loss is formulated as

$$\overline{u}_c = \sum_{u,v} u \cdot \mathcal{P}_{ego}^{c,u,v}/z_k, \ \ \overline{v}_c = \sum_{u,v} v \cdot \mathcal{P}_{ego}^{c,u,v}/z_k, \qquad (6)$$

$$\mathcal{L}_c = \sum_c \sum_{u,v} \|\langle u, v\rangle - \langle \overline{u}_c, \overline{v}_c\rangle\| \cdot \mathcal{P}_{ego}^{c,u,v}/z_c, \qquad (7)$$

where $\overline{u}_c$ and $\overline{v}_c$ represents the center of the c-th localization map along axis u, v, and $z_c = \sum_{u,v} \mathcal{P}_{ego}^{c,u,v}$ is a normalization term. The concentration loss forces the high activation regions of the localization maps to be close to the geometric center.

Overall, we train the whole framework in an end-to-end manner, and use the following loss to optimize the model:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{cos}\mathcal{L}_{cos} + \lambda_c\mathcal{L}_c, \qquad (8)$$

where λ_{cos}, and λ_c are loss weights that balance the three terms. $\mathcal{L}_{cls}$ stands for the cross-entropy losses from the two branches. At test time, only the ego branch is maintained, taking an affordance label t and an egocentric image as input, the network extracts the t-th localization map as the prediction of affordance region.

4. Experiments

4.1. Experimental Setting

Dataset and Metrics. We evaluate our method in the Affordance Grounding Dataset (AGD20K) [33], which is the only large-scale image dataset with both exocentric and egocentric views. AGD20K is comprised of 20,061 exocentric images and 3,755 egocentric images, and is annotated with 36 commonly used affordances. Following prior affordance grounding work [14,38], the ground truth of this dataset initially consists of densely annotated points in corresponding affordance regions, and a Gaussian blur is then applied over each point to get final heatmaps. Moreover, AGD20K can be evaluated in two different settings: 1) In the seen setting, object categories in training and test sets are identical. 2) In the unseen setting, there is no object category intersection between training and test sets, e.g., the

State-of-the-Art from Relevant Tasks		Seen			Unseen		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Weakly Supervised Object Localization*	EIL [34]	1.931	0.285	0.522	2.167	0.227	0.330
	SPA [40]	5.528	0.221	0.357	7.425	0.169	0.262
	TS-CAM [16]	1.842	0.260	0.336	2.104	0.201	0.151
Weakly Supervised Affordance Grounding	Hotspots [38]	1.773	0.278	0.615	1.994	0.237	0.577
	Cross-view-AG [33]	1.538	0.334	0.927	1.787	0.285	0.829
	Cross-view-AG+ [32]	1.489	0.342	0.981	1.765	0.279	0.882
	AffCorrs† [18]	<u>1.407</u>	<u>0.359</u>	<u>1.026</u>	<u>1.618</u>	<u>0.348</u>	<u>1.021</u>
	LOCATE (Ours)	**1.226**	**0.401**	**1.177**	**1.405**	**0.372**	**1.157**

Table 1. Comparison to state-of-the-arts from relevant tasks on AGD20K dataset. Results of * are taken from [33], and † denotes the adapted AffCorrs. The **best** and <u>second-best</u> results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better).

model observes how humans hold a hammer and anticipates where to hold a knife.

As for the metrics, referring to previous affordance grounding work [14, 30, 33, 38], we adopt the commonly used Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) to evaluate the similarity and correspondence of distributions between ground truth and prediction. Detailed calculation of each metric is shown in the supplementary material.

Implementation Details. We use the ImageNet [11] pretrained (without supervision) DINO-ViT-S [6] with patch size 16 to generate deep features. In each iteration, N exocentric images along with one egocentric image are taken as input (N is set to 3). Images are first resized to 256×256 and then randomly cropped to 224×224 followed by random horizontal flipping. SGD with learning rate 1e-3, weight decay 5e-4, and batch size 16 is used for parameter optimization. Loss weight coefficients (λ_{cos}, λ_c) are set to (1, 0.07), and the margin α is set to 0.5. For the first epoch, we warm up the network without $\mathcal{L}_{cos}$, as initial localization maps are not accurate for supervision.

4.2. Comparison to State-of-the-Art

To conduct a comprehensive comparison, we also display the results of state-of-the-art methods from a relevant task, i.e., weakly supervised object localization. As shown in Table 1, in both seen and unseen settings, LOCATE outperforms all other methods with a considerable margin on all metrics. In particular, compared to the state-of-the-art affordance grounding method Cross-view-AG+ [32], we improve the KLD by 20.4%, SIM by 33.3%, and NSS by 31.2% in the unseen setting. Cross-view-AG+ is an extended version of Cross-view-AG, but still performs the knowledge transfer based on global pooled embeddings at the image level, thus bringing only minor improvement. AffCorrs [18] is a method that focuses on one-shot part affordance grounding, and it also uses the pretrained DINO-ViT

Methods	Params (M)	Time (s)
EIL [34]	42.41	0.019
SPA [40]	69.28	0.081
TS-CAM [16]	85.86	0.023
Hotspots [38]	132.64	0.087
Cross-view-AG [33]	120.03	0.023
Cross-view-AG+ [32]	82.27	0.022
AffCorrs† [18]	**6.50**	0.205
LOCATE (Ours)	**6.50**	**0.011**

Table 2. Comparison of learnable parameters and inference time. The inference time is evaluated on a 3090Ti GPU. † denotes the adapted AffCorrs.

features to do part matching. However, AffCorrs needs a pixel-level mask as a query, and there is no domain gap during the knowledge transfer. To make AffCorrs comparable in our problem setting, we adapt its structure by replacing the query annotated mask with our CAM estimator. The results verify that AffCorrs can also achieve good performance, but still considerably inferior to LOCATE.

In Table 2, we make comparisons in terms of model parameters and inference time. Since our framework is built on a frozen small-sized vision transformer (ViT-small), the training process is efficient with a small number of parameters. For example, LOCATE only has 5.4% of learnable parameters in Cross-view-AG. Additionally, we use a large patch size 16 for the vision transformer, which constrains the input sequence length and greatly reduces computation cost. Therefore, the inference time of LOCATE is also faster than most other methods. By contrast, the adapted AffCorrs runs much slower than LOCATE, as it incorporates an additional CRF post-processing step.

We further visualize the qualitative comparisons with state-of-the-art affordance grounding methods. As shown in Fig. 5, we compare our results with Hotspots [38], Cross-
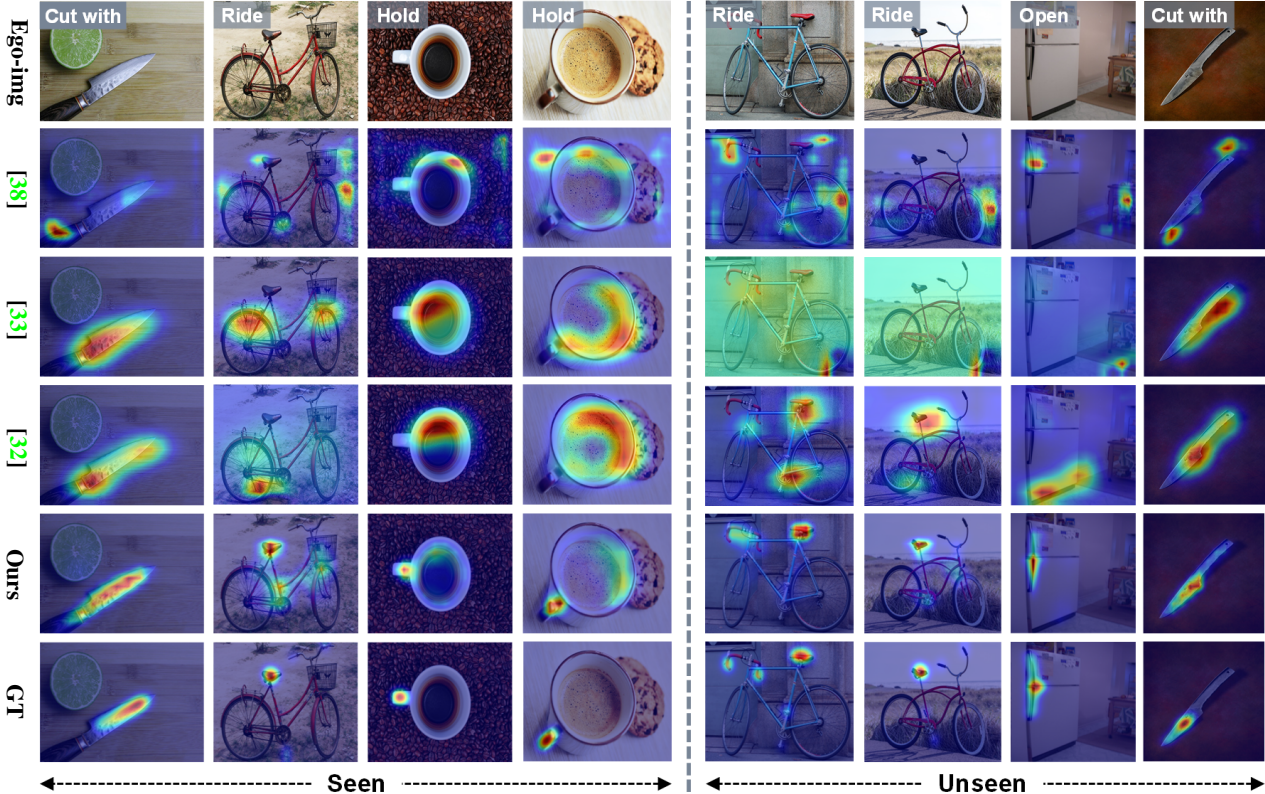
Figure 5. Qualitative comparison between our approach and state-of-the-art affordance grounding methods (Hotspots [38], Cross-view-AG [33], and Cross-view-AG+ [32]). For the unseen setting, the displayed objects are not in the training set. For example, the model learns where a *motorcycle* can be ridden in training, and locates rideable area for the *bicycle* at test time.

view-AG [33] and Cross-view-AG+ [32]. We observe that the proposed LOCATE can make more concentrated and accurate predictions. Especially for complex objects like bicycles and refrigerators, even in the unseen setting, our method can still locate the saddle of bicycles for riding, and the handle of fridges for opening. In comparison, the results of Cross-view-AG for bicycles are quite noisy. More visualization results are in the supplementary material.

4.3. Ablation Study

Knowledge Transfer Manner. We first investigate the impact of knowledge transfer manner. Previous affordance grounding methods [33, 38] simply pull close the global embeddings (produced by global average pooling) of two branches to perform global knowledge transfer (GKT). In contrast, we set up an experiment to implement regional knowledge transfer (RKT), which generates the embeddings via masked average pooling between CAM-produced localization maps and feature maps. The results are shown in Table 3, regional knowledge transfer (RKT) outperforms global knowledge transfer (GKT) on all metrics, demonstrating the effectiveness of filtering irrelevant information.

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
GKT	1.732	0.267	0.810	1.971	0.221	0.626
RKT	1.516	0.320	1.074	1.823	0.259	0.850
+ $\mathcal{L}_c$	1.491	0.326	1.091	1.750	0.274	0.948
+ $\mathcal{S}$	1.236	0.397	1.178	1.439	0.358	1.130
+ $\mathcal{S} + \mathcal{L}_c$	1.226	0.401	1.177	1.405	0.372	1.157

Table 3. Ablation results of the proposed LOCATE framework. GKT/RKT means global/regional knowledge transfer. $\mathcal{S}$ denotes PartSelect combined with $\mathcal{L}_{cos}$, and $\mathcal{L}_c$ is the concentration loss.

PartSelect and Concentration Loss. Based on the regional knowledge transfer, we analyze the effect of PartSelect and concentration loss. As shown in Table 3, directly applying the concentration loss $\mathcal{L}_c$ can only bring marginal improvement. The reason is $\mathcal{L}_c$ can make egocentric predictions more concentrated, but fail to guide it to focus on the right affordance area. Nonetheless, when adding PartSelect and using cosine embedding loss $\mathcal{L}_{cos}$ as explicit supervision, the performance is greatly boosted, which proves the effectiveness of the part-level knowledge transfer scheme. In ad-
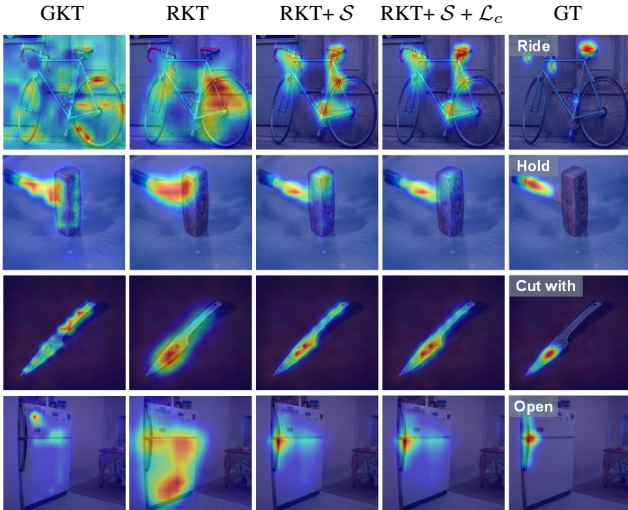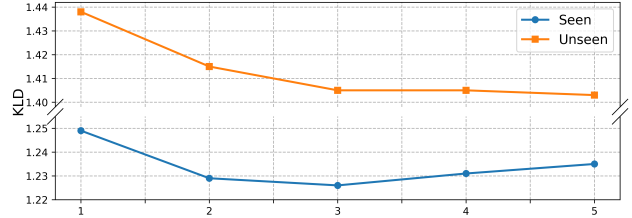
Figure 6. Visualization of the qualitative improvements.

(a) Number of prototypes N

(b) Number of exocentric images K

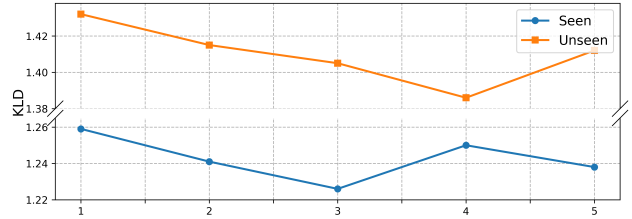Figure 7. Ablation study on the number of prototypes and exocentric images (lower is better).

dition, to check the qualitative improvement, we visualize the affordance grounding results in Fig. 6. It is clear that GKT tends to locate the wrong affordance area, while RKT can sometimes find the right region, but only give coarse grounding results. After adding PartSelect, the results become much more part-focused, and the concentration loss further makes the grounding maps more robust.

Number of Prototypes/Exocentric Images. We then explore the impact from the number of prototypes K and exocentric images N. From Fig. 7a, we observe that the model yields the best performance with three prototypes in seen setting, which is consistent with our statement that interaction areas typically consist of human, object part, and background information. While for the unseen setting, $N = 5$ achieves the best results, but improvement is minor. One reason lies in that more prototypes segment objects into more small parts, which boosts the generalization ability. As for the number of exocentric images, we find that more exocentric images can alleviate the impact of interaction diversity and occlusion, thus providing more robust knowledge for the egocentric branch. As shown in Fig. 7b, the model gets largely improved when increasing the number of exocentric images from 1 to 3 in both seen and unseen settings. Finally, we set both N and K to 3.

Different Feature Extractors. In LOCATE, we employ pretrained DINO features based on ViT, which have been proven to encode high-level semantic information [1]. To investigate the impact of different backbones, we conducted experiments with DINO features trained on the ResNet-50 [22]. From the results in Table 4, we observe that using DINO-ViT features directly (without the proposed PartSelect) can only obtain similar or even inferior results to their ResNet counterpart. After incorporating PartSelect, the results of both backbone features can be improved under the seen setting, but ViT features show better potential in en-

	RN-50	ViT-S/16	PartSelect	KLD↓	SIM↑	NSS↑
Seen	✓			1.482	0.334	1.005
		✓		1.491	0.326	1.091
	✓		✓	1.449	0.340	1.021
		✓	✓	1.226	0.401	1.177
Unseen	✓			1.701	0.287	0.962
		✓		1.750	0.274	0.948
	✓		✓	1.707	0.287	0.949
		✓	✓	1.405	0.372	1.157

Table 4. Ablation study on different feature extractors.

hancing the performance due to their part-aware property. In the unseen setting, PartSelect does not yield improvement for ResNet-based features, while ViT features obtain consistent gains.

5. Conclusion

In this paper, we propose a framework named LOCATE to address the weakly supervised affordance grounding task by observing human-object interaction images. Specifically, we first localize where the interaction happens for the exocentric interactions, and then design a module called PartSelect to pick the affordance-specific information from the interaction regions. Finally, we transfer the learned knowledge to the egocentric view to perform affordance grounding with only image-level affordance labels. The proposed LOCATE achieves state-of-the-art results with far fewer parameters and faster inference speed.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For*, 2022. 3, 8

[2] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020. 1

[3] Paola Ardón, Eric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019. 1

[4] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pretraining of image transformers. *ICLR*, 2022. 3

[5] Christopher J Burke, Philippe N Tobler, Michelle Baddeley, and Wolfram Schultz. Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436, 2010. 1

[6] 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. 2, 3, 6

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018. 3

[8] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, 2022. 3

[9] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *CVPR*, 2022. 3

[10] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 1, 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[12] Thanh Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. *ICRA*, 2018. 1, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[14] Kuan Fang, Te Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2Vec: Reasoning Object Affordances from Online Videos. *CVPR*, 2018. 1, 3, 5, 6

[15] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *ArXiv*, abs/2206.06363, 2022. 3

[16] 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. 3, 6

[17] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979. 1

[18] Denis Hadjivelichkov, Sicelukwanda Zwane, Marc Deisenroth, Lourdes Agapito, and Dimitrios Kanoulas. One-Shot Transfer of Affordance Regions? AffCorrs! *CoRL*, 2022. 3, 6

[19] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *ICLR*, 2021. 3

[20] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 1

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[24] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, 2019. 5

[25] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *CVPR*, 2022. 3

[26] Mia Kokic, Johannes A Stork, Joshua A Haustein, and Danica Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE, 2017. 1

[27] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 2

[28] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *CVPR*, 2021. 3

[29] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 3

[30] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 6

[31] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *arXiv preprint arXiv:2108.05675*, 2021. 3

[32] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *arXiv preprint arXiv:2208.13196*, 2022. 1, 6, 7

[33] 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, 3, 5, 6, 7

[34] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *CVPR*, 2020. 3, 6

[35] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, 2022. 3

[36] Austin Myers, Angjoo Kanazawa, Cornelia Fermuller, and Yiannis Aloimonos. Affordance of Object Parts from Geometric Features. *Int. Conf. Robot. Autom.*, pages 5–6, 2015. 2

[37] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 1, 2

[38] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. *ICCV*, 2019. 1, 2, 3, 5, 6, 7

[39] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 1, 2

[40] 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. 3, 6

[41] Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *ICCVW*, 2017. 2

[42] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. *CVPR*, 2017. 2

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3

[44] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 3

[45] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 3

[46] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, 2022. 3

[47] Huadong Wu, Zhanpeng Zhang, Hui Cheng, Kai Yang, Jiaming Liu, and Ziying Guo. Learning affordance space in physical world for vision-based robotic object manipulation. In *ICRA*, 2020. 1

[48] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *CVPR*, 2022. 3

[49] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *CVPR*, 2022. 3

[50] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 3

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2, 3

[52] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, 2022. 3