

Selective Contrastive Learning for Weakly Supervised Affordance Grounding

WonJun Moon[†] Hyun Seok Seong[†] Jae-Pil Heo^{*}
Sungkyunkwan University

{wjun0830, gustjrd195, jaepilheo}@skku.edu

Abstract

Facilitating an entity’s interaction with objects requires accurately identifying parts that afford specific actions. Weakly supervised affordance grounding (WSAG) seeks to imitate human learning from third-person demonstrations, where humans intuitively grasp functional parts without needing pixel-level annotations. To achieve this, grounding is typically learned using a shared classifier across images from different perspectives, along with distillation strategies incorporating part discovery process. However, since affordance-relevant parts are not always easily distinguishable, models primarily rely on classification, often focusing on common class-specific patterns that are unrelated to affordance. To address this limitation, we move beyond isolated part-level learning by introducing selective prototypical and pixel-contrastive objectives that adaptively learn affordance-relevant cues at both the part and object levels, depending on the granularity of the available information. Initially, we find the action-associated objects in both egocentric (object-focused) and exocentric (third-person example) images by leveraging CLIP. Then, by cross-referencing the discovered objects of complementary views, we excavate the precise part-level affordance clues in each perspective. By consistently learning to distinguish affordance-relevant regions from affordance-irrelevant background context, our approach effectively shifts activation from irrelevant areas toward meaningful affordance cues. Experimental results demonstrate the effectiveness of our method.

1. Introduction

Humans learn to interact with objects by observing others and recognizing relevant object parts in interactions [2, 23]. Similarly, weakly supervised affordance grounding focuses on identifying which parts of an object afford particular interactions within the environment in which humans typically

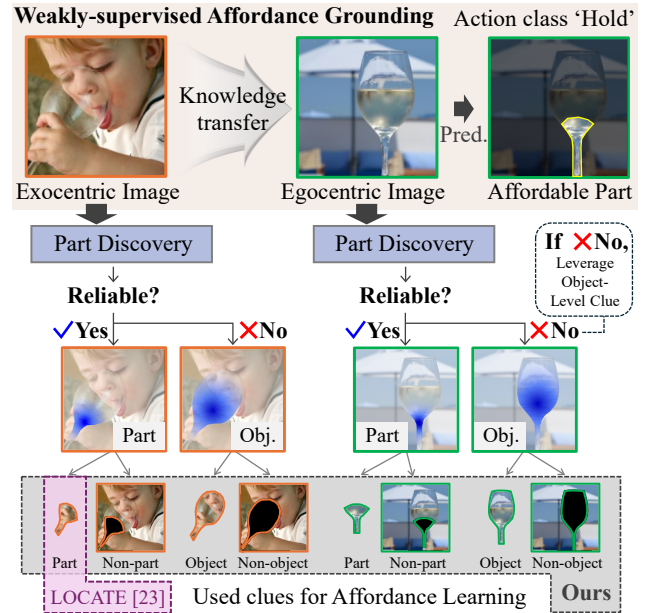


Figure 1. (Up) Goal of WSAG is to identify action-affordable parts within the egocentric image, given exocentric images as contextual hints. (Down) To perform affordance learning, we first discover the part-relevant clues from both egocentric and exocentric images. When these parts are deemed reliable in representing affordance-relevant regions, the model learns to distinguish these parts from the other parts. If not, we instead utilize object-level clues to distinguish objects from the background. Compared to our baseline (LOCATE [23]), which only exploits reliable parts of exocentric images, our approach extends affordance learning to learn from both egocentric and exocentric views and also both from affordance-relevant and affordance-irrelevant clues. By leveraging all these types of clues within the mini-batch at once, the model learns to distinguish affordance-relevant parts from representations of other affordance classes and backgrounds.

learn [17, 23, 28, 34, 37, 38, 47]. Specifically, a target egocentric image (object-focused) is provided with an action class name and a few exocentric images (human-object interaction examples given in third-person view) to localize the affordable parts within egocentric image [28, 34, 38]. Then, the model is trained to localize action-affordable parts when an egocentric image is provided along with an action class.

[†] Equal contribution

^{*} Corresponding author

In this vein, knowledge distillation is widely studied [23, 47], along with action classification to produce a class activation map (CAM) [54] for localization. For example, LOCATE [23] introduced a part-level distillation approach. It extracts action-affordable parts from exocentric images by segmenting interaction-involved regions identified by CAM. These action-affordable parts are distilled into the egocentric image representations only when they are precisely identified, enabling alignment with affordance-relevant regions.

Yet, as training without dense annotation progresses, they tend to locate distinguishable parts necessary for action classification even if they are not directly related to the affordable part. This is because affordance-relevant clues are not always clearly distinguishable, thus, the distillation is only applied intermittently. To address this, we go beyond solely focusing on part feature distillation; our primary objective is to consistently provide contextual cues to distinguish between affordance-relevant and affordance-irrelevant representations. The overall intuition is illustrated in Fig. 1.

We begin by collecting object-level affordance-relevant clues from both egocentric and exocentric images, then gradually refine them to part-level clues. The model is then trained to focus on these affordable parts via dedicated selective contrastive learning. Specifically, if the identified part clue is deemed to correspond to an affordance region reliably, the model learns to distinguish it from other irrelevant parts. Conversely, if the identified part is deemed unsuitable, the model is trained to distinguish the target object clue (identified using the object affinity map) from the background, preventing attention to affordance-irrelevant regions.

First, to collect the clues for action-associated objects, we leverage CLIP [39] to generate an object affinity map that encompasses affordance-relevant parts. The identified target object then serves as a basis for discovering part-level affordance clues. For part discovery within the exocentric view, we refine the part discovery algorithm from LOCATE [23] by leveraging the target object to improve precision. Specifically, the object affinity map is used to filter out object-irrelevant part candidates, ensuring that the affordance-relevant parts belong to the target object. Conversely, to extract part clues in the egocentric view, we exploit the properties of foundation models that CLIP tends to be more responsive to prominent objects [9]. Specifically, we assess part cues by analyzing the difference in model activation between egocentric and exocentric images, where responses tend to be weaker in exocentric images due to smaller object scales and occlusions.

Upon gathering object and part clues from both views, we design two types of contrastive learning to leverage the collected affordance-relevant clues. Initially, we propose prototypical contrastive learning to exploit affordance-relevant clues from the exocentric view, offering several key advantages over previously used pairwise distillation

strategies [23, 47]. While pairwise distillation focuses solely on reducing the distance between representations of paired egocentric and exocentric images, prototypical contrastive learning not only encourages egocentric-exocentric aligned representations but also distinguishes each prototype from diverse background information and the prototypes of other action classes. This enables the model to capture more discriminative representations specific to each action class. On the other hand, pixel-level contrastive learning further optimizes the localization of affordable parts with precise pixel-level clues. Specifically, it directly uses affordance-relevant clues in egocentric images to disentangle affordance-relevant pixels from the others in each image. This facilitates pixel representations to be distinguished based on their affordance relevance at the level of gathered clues.

To sum up, our contributions are: (i) We propose prototypical contrastive learning to benefit part representation learning by leveraging the semantics of other action classes and backgrounds. (ii) We propose pixel contrastive learning to supplement the fine-grained localization of affordance-relevant regions. (iii) We present a post-processing step to calibrate CAM prediction by leveraging CLIP’s capability to detect text-specified objects. Our approach consistently outperforms (iv) Our approach demonstrates superior performance over prior methods, particularly in challenging unseen scenarios that closely reflect real-world conditions.

2. Related Work

2.1. Visual Affordance Grounding

Visual affordance grounding aims to locate the responsible object parts to certain actions [24]. To minimize the gap between perception and action, extensive attention is being put into affordance grounding among the researchers of computer vision and robotics [14, 20, 23]. Initially studied in a supervised setting [11, 33, 35], affordance grounding is recently being studied more in weakly supervised scenarios where costly dense annotations are not required [6, 17, 28, 38, 47]. For example, LOCATE [23] uses CAM to identify interaction-involved regions and applies K-means clustering to find the affordance-relevant parts in exocentric images for distillation. WSMA [47] exploited the semantics of CLIP through an attention mechanism to address the limitation of discrete classification labels in illustrating the semantics of actions. Also, more recent works utilize diverse foundation models, such as ALBEF [26], SAM [19], LLaVA [27], and GPT [1], to obtain part-level knowledge [6, 17, 38, 40]. Our approach, despite not relying on recent foundation models, significantly outperforms them by effectively addressing cases where reliable parts cannot be identified and leveraging background context to prevent the model from focusing on affordance-irrelevant regions.

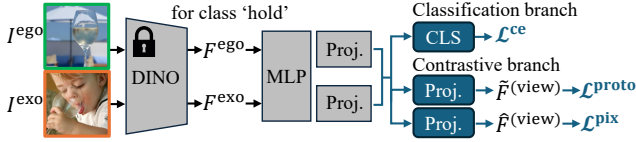


Figure 2. Overall flow. Egocentric and exocentric images are processed to perform classification and selective contrastive learning. Note that $(view) \in \{ego, exo\}$.

2.2. Weakly Supervised Object Localization

Weakly Supervised Object Localization (WSOL) aims to localize objects using only image-level labels. Conventionally, CAM-based methods have been widely studied due to their effectiveness [10, 30, 46, 48, 49, 51, 53]. Yet, these often suffer from shortcut learning [13], which limits CAM coverage, making CAM expansion a common strategy for WSOL. For example, HaS [21] randomly masks image patches during training, CutMix [50] enhances masked images, and LoRot [32] introduces pretext tasks involving random scaling and positioning to broaden the model’s receptive field. A similar challenge arises in WSAG, where the goal is to localize affordance-relevant parts for a given action class, independent of object categories. Although WSAG also struggles with the model focusing on commonly appearing details within each action class, CAM expansion is not always a suitable solution, as affordance-relevant parts are often small. To address this, we propose a selective strategy that adaptively determines whether to expand the CAM to the object region when a reliable part cannot be identified or to concentrate CAM activation when a reliable part is available.

2.3. Contrastive Learning

Contrastive learning pulls together instances with positive relationships while pushing apart those with negative relationships [4, 7, 16, 31]. It has been employed in various fields by adapting the criteria for determining the relationships between instances. For instance, augmented pairs of the same instance are regarded as positives in an unsupervised setting [7], while samples within the same class are treated as positives in a supervised setting [18]. For WSAG, LLM has been employed to derive the relationships between interaction types [17]. Also, there is significant variation in the units to which contrastive learning is applied. For example, while the images are typical units [7, 8], prototypes [25], pixels [41–44, 52], or even the similarity between modalities [39] are popular sources. In this work, we introduce prototypical and pixel contrastive learning that adaptively selects the training level to optimize both object- and part-level regions.

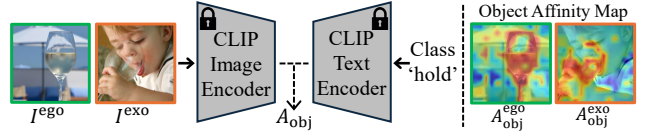


Figure 3. Illustration of object discovery. The object affinity map is derived from CLIP as a zero-shot image-text similarity map.

3. Method

3.1. Method Overview

In Fig. 2, we illustrate the overall framework. Given a pair of an egocentric image I^{ego} and multiple exocentric images I^{exo} , the inputs are processed using DINO [5] followed by projection layers. Then, for prototypical and pixel contrastive learning, features are further projected to obtain $\tilde{F}^{(view)}$ and $\hat{F}^{(view)}$, respectively, where $(view) \in \{ego, exo\}$. We note that the features of the egocentric image are represented as $\tilde{F}^{ego}, \hat{F}^{ego} \in \mathbb{R}^{B \times H \times W \times D}$, while the features of the exocentric images are given by $\tilde{F}^{exo}, \hat{F}^{exo} \in \mathbb{R}^{B \times E \times H \times W \times D}$, where B denotes the batch size, H and W represent the spatial dimensions, D is the feature dimension, and E indicates the number of exocentric images. While the contrastive learning branch focuses on learning affordance knowledge within the egocentric images, the classification branch with the shared classifier captures the shared semantic information between egocentric and exocentric views. For inference, CAM C^{ego} is derived from classification branch using only egocentric images and affordance text prompts.

To conduct selective contrastive learning, we first establish target supervision by identifying action-associated objects in Sec. 3.2. Subsequently, in Sec. 3.3 and Sec. 3.4, we introduce prototypical and pixel contrastive learning, respectively, along with the part-level target discovery process.

3.2. Object Discovery

As illustrated in Fig. 3, we leverage CLIP to define the object affinity map. Particularly, we employ the strategy of ClearCLIP [22] to enhance local discriminability in visual features. Given egocentric features and exocentric features from CLIP visual encoder, we calculate cosine similarity with CLIP text features of action prompt to obtain an object affinity map for each perspective, namely $A_{obj}^{ego} \in \mathbb{R}^{B \times H \times W}$ and $A_{obj}^{exo} \in \mathbb{R}^{B \times E \times H \times W}$ (Details for action prompts are in Appendix). Note that the term *object affinity map* is derived from its characteristic to highlight affordance-relevant objects when action prompt is given, as shown in Fig. 7.

3.3. Prototypical Contrastive Learning

Prototypical contrastive learning operates upon gathered affordance-relevant clues within exocentric view. Simply put, prototypes for affordable parts in exocentric images are distilled towards corresponding prototypes within egocentric

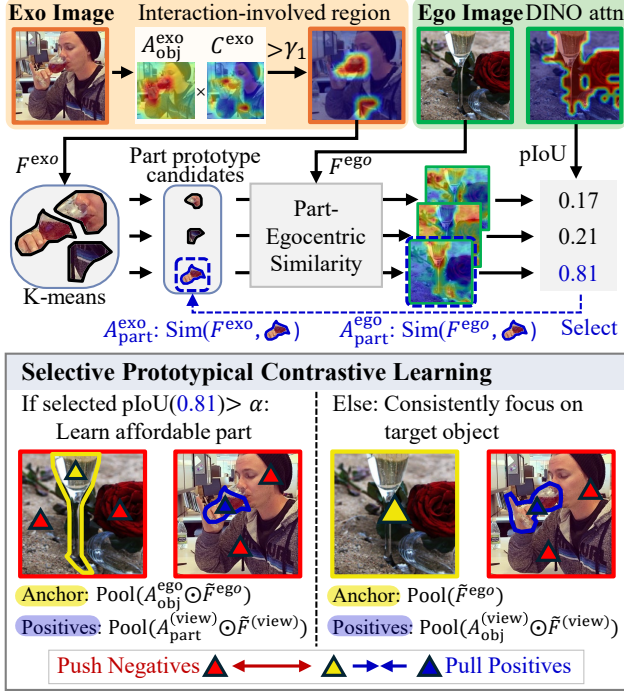


Figure 4. Illustration of prototypical contrastive learning. (Up) Process of identifying part clues in exocentric images. Discovered objects are segmented to extract part candidates, which are then matched with DINO’s attention map. (Down) Prototypical contrastive learning is selectively applied based on the reliability of part clues. When reliable, object anchors in egocentric images are attracted toward part clues, but otherwise image anchors are drawn toward object clues in exocentric images.

images via contrastive learning. The key advantage of prototypical contrastive approach over previous works is that it encompasses the process of learning negative relationships between prototypes. Thus, we claim that the classification bias towards affordance-irrelevant regions, *i.e.*, background context, and non-affordable object parts, can be mitigated.

Part-level Clues within Exocentric Images. We begin by illustrating the process of gathering part-level clues in exocentric images, as shown in Fig. 4. Specifically, we adapt the algorithm used from previous work [23]. To illustrate, whereas previous work directly thresholded CAM prediction C^{exo} to identify interaction-involved regions in exocentric images, we first combined C^{exo} with the object affinity map A_{obj}^{exo} before applying a threshold γ_1 . This ensures that the region of interest is constrained to object regions, mitigating the risk of imprecise CAM predictions and improving the affordance relevance of the extracted affordance cues. The rest of the process follows that of previous work [23]. First, based on the intuition that CAM regions consist of background, affordance-relevant part, and other elements, K-means clustering (K=3) is applied. Then, the centroids (candidates of part prototype) are compared with the egocentric DINO [5] feature F^{ego} to generate part-egocentric similarity maps.

These maps are then assessed to determine whether each corresponding centroid represents the affordance-relevant parts by comparing them with the self-attention map of the egocentric image from DINO [5], measured by pIoU [23] (DINO attention map can be replaced by object affinity map). Finally, only the centroid corresponding to the highest pIoU that exceeds a threshold α is selected as the designated part. If either condition is not met, the part (centroid) is considered unreliable and excluded from training. Consequently, for instances with reliable part prototype, part affinity maps $A_{part}^{(view)}$ are defined as the similarity between the selected part prototype and spatial features (*i.e.*, F^{ego} and F^{exo}).

Selective Prototypical Contrastive Learning. Due to the inconsistent availability of affordable part clues, the typical approach is to exploit the knowledge in exocentric images only when the reliable part is discovered. Yet, this triggers the affordance grounding task to become heavily reliant on classification tasks, which is vulnerable in capturing target object parts since its goal is to find the most discriminative features for action classification.

Therefore, we design a loss function that consistently leverages the knowledge of interaction-involved regions in exocentric images throughout the training. Specifically, our prototypical contrastive learning integrates learning-level selectivity for both the target and anchor representations. When the discovered part prototype within exocentric images is deemed reliable, we use it as the target prototype for distillation into the egocentric object prototype. Otherwise, we define the object prototype to serve as the target and the entire egocentric image as an anchor. This design, which sets the object prototype as the default distillation target, encourages the model to consistently focus on the target object while disregarding background context in egocentric images. Furthermore, when part supervision is available, it reinforces attention to affordable parts, enhancing the model’s ability to capture details of affordable parts.

To leverage the object/part clues in prototypical contrastive learning, we initially construct prototypes. Particularly, four types of prototypes, namely P^{ego+} , P^{ego-} , P^{exo+} and P^{exo-} , are produced which refer to the positive and negative prototypes of object/part clues in each view depending on the level of gathered clues. In particular, these positive and negative prototypes are constructed with following functions (Φ^+ and Φ^-) using instance feature $Z \in \mathbb{R}^{H \times W \times D}$, target clue $M \in \mathbb{R}^{H \times W}$ and CAM prediction $C \in \mathbb{R}^{H \times W}$:

$$\begin{aligned} \Phi^+(Z, M) &= \text{norm}(\text{Pool}(Z \odot M)), \\ \Phi^-(Z, M, C) &= \text{norm}(\text{Pool}(Z \odot (\beta - M \odot C))), \end{aligned} \quad (1)$$

where $\text{norm}(\cdot)$ indicates Frobenius normalization along channel axis, $\text{Pool}(\cdot)$ denotes spatial average pooling, and β is a bias term to prevent training instability incurred by imprecise CAM C at the initial training epoch. Note that \odot is defined as $(\mathbf{X} \odot \mathbf{Y})_{i,j,k} = (x_{i,j,k}) \times (y_{i,j})$, $\forall i \in \{1, \dots, H\}, \forall j \in$

$\{1, \dots, W\}, \forall k \in \{1, \dots, D\}$ to apply Hadamard product between \mathbf{X} and \mathbf{Y} in different shapes. In short, the positive prototype maintain a consistent focus on target regions by masking with target clue M , which is often more precise than CAM prediction, while the background prototype captures general background semantics and unaffordable parts.

Subsequently, let \mathbb{I} denote the index set of both the exocentric and egocentric instances within the mini-batch which represents instances with precise part-level prototypes (we assume that there is only one exocentric image per egocentric image in this subsection, thereby \mathbb{I} can be shared for simplicity). Then, the egocentric anchor z_b^{ego} of b -th instance and the prototypes are formed as follows:

$$z_b^{\text{ego}} = \begin{cases} \Phi^+(\tilde{F}_b^{\text{ego}}, A_{\text{obj},b}^{\text{ego}}) & \text{if } b \in \mathbb{I}, \\ \text{norm}(\text{Pool}(\tilde{F}_b^{\text{ego}})) & \text{otherwise,} \end{cases}$$

$$P_b^{(\text{view})+} = \begin{cases} \Phi^+(\tilde{F}_b^{(\text{view})}, A_{\text{part},b}^{(\text{view})}) & \text{if } b \in \mathbb{I}, \\ \Phi^+(\tilde{F}_b^{(\text{view})}, A_{\text{obj},b}^{(\text{view})}) & \text{otherwise,} \end{cases} \quad (2)$$

$$P_b^{(\text{view})-} = \begin{cases} \Phi^-(\tilde{F}_b^{(\text{view})}, A_{\text{part},b}^{(\text{view})}, C^{(\text{view})}) & \text{if } b \in \mathbb{I}, \\ \Phi^-(\tilde{F}_b^{(\text{view})}, A_{\text{obj},b}^{(\text{view})}, C^{(\text{view})}) & \text{otherwise.} \end{cases}$$

Consequently, our selective prototypical contrastive learning for b -th instance in mini-batch is expressed as:

$$\mathcal{L}_b^{\text{proto}} = \frac{-1}{|\mathbf{P}_b^+|} \sum_{p \in \mathbf{P}_b^+} \log \frac{\exp(z_b^{\text{ego}} \circ p / \tau)}{\sum_{n \in (\mathbf{P}_b^+ \cup \mathbf{P}_b^-)} \exp(z \circ n / \tau)}, \quad (3)$$

where \circ and τ denote dot product and temperature parameter, respectively. \mathbf{P}_b^+ and \mathbf{P}_b^- which represent the sets of positive and negative prototypes for b -th instance are defined as:

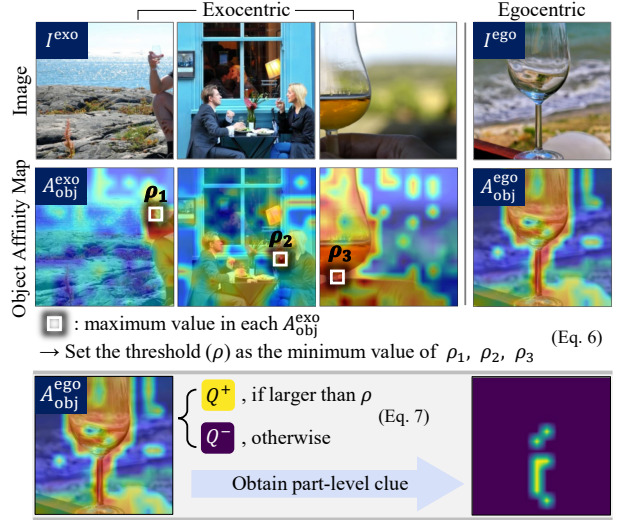
$$\mathbf{P}_b^+ = \bigcup_{(\text{view})} \bigcup_{i \in \mathcal{B}} \{P_i^{(\text{view})+} | \delta(P_i^{(\text{view})+}) = \delta(z_b^{\text{ego}})\}, \quad (4)$$

$$\mathbf{P}_b^- = \bigcup_{(\text{view})} \left\{ \bigcup_{i \in \mathcal{B}} \{P_i^{(\text{view})-} | \delta(P_i^{(\text{view})-}) = \delta(z_b^{\text{ego}})\} \right. \\ \left. \bigcup_{j \in \mathcal{B}} \{P_j^{(\text{view})+} | \delta(P_j^{(\text{view})+}) \neq \delta(z_b^{\text{ego}})\} \right\}, \quad (5)$$

where i, j denote the index from batch index set \mathcal{B} , and δ is a function to output the action class label of given instances. Consequently, prototypical contrastive learning directs the model's activation toward affordance-relevant regions. Specifically, object-level learning enhances focus on object regions, and when affordance-relevant parts are present, it further refines features to capture part-specific information within the object.

3.4. Pixel Contrastive Learning

In prototypical contrastive learning, we encourage the model to prioritize foreground objects over the entire image and,



Acquiring part-level knowledge of class 'hold' for egocentric image
Figure 5. An illustration of binarizing objects within egocentric images based on affordance criterion. The most salient pixel in each exocentric object affinity map serves as a reference, establishing a criterion to classify each pixel in the egocentric image as part of an affordable region (Q^+) or a non-affordable region (Q^-). The minimum value among ρ_1, ρ_2 , and ρ_3 is used as criterion.

within these objects, to focus on specific parts. However, we remark that only the implicit guidance is provided on each pixel of affordable parts. Thus, we additionally propose pixel contrastive learning to supplement the fine-grained localization capability by learning correspondences between pixels in each egocentric image.

Part-level Clues within Egocentric Images. Symmetrically to the use of egocentric view for gathering part clues in exocentric images, we utilize exocentric view as contextual cues to capture part clues in egocentric images. Specifically, we leverage the property of foundation models (CLIP) that they are more responsive to salient objects [9]. Thus, we expect stronger activations for affordable parts in egocentric images compared to their exocentric counterparts when matched with text prompts describing the part to perform a specific action with. This is because exocentric images depict objects in use, often capturing them at a small scale and making them more susceptible to occlusions.

The overall process for egocentric part discovery is illustrated in Fig. 5. Initially, we determine the criterion ρ which is used to distinguish pixels that belong to affordable parts in egocentric images. The logic for deriving $\rho \in \mathbb{R}^B$ is:

$$\rho = \min_{e \in E} \max_{h, w \in H, W} A_{\text{obj}}^{\text{exo}}. \quad (6)$$

To clarify, we first compute the maximum value across the spatial dimensions ($H \times W$) for exocentric object affinity map $A_{\text{obj}}^{\text{exo}} \in \mathbb{R}^{B \times E \times H \times W}$, resulting in a tensor of shape $B \times E$. Then, we select the minimum value along the E axis, which is the number of exocentric images paired with each

Table 1. Performance comparison on the AGD20K and HICO-IIF datasets.

Method	Model	AGD20K-Seen			AGD20K-Unseen			HICO-IIF		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Zero-Shot Vision-Language Model										
Clear-CLIP [22]	CLIP	1.573	0.294	0.945	1.723	0.262	0.976	1.746	0.252	1.032
Weakly Supervised Object Localization										
SPA [36]	-	5.528	0.221	0.357	7.425	0.169	0.262	-	-	-
EIL [30]	-	1.931	0.285	0.522	2.167	0.277	0.330	-	-	-
TS-CAM [12]	DeiT	1.842	0.260	0.336	2.104	0.201	0.151	-	-	-
Weakly Supervised Affordance Grounding										
Hotspots [34]	ResNet50	1.773	0.278	0.615	1.994	0.237	0.577	-	-	-
Cross-view-AG [28]	ResNet50	1.538	0.334	0.927	1.787	0.285	0.829	1.779	0.263	0.946
Cross-view-AG+ [29]	ResNet50	1.489	0.342	0.981	1.765	0.279	0.882	1.836	0.256	0.883
LOCATE [23]	DINO	1.226	0.401	1.177	1.405	0.372	1.157	1.593	0.327	0.966
WSMA [47]	DINO+CLIP	1.176	0.416	1.247	1.335	0.382	1.220	1.465	0.358	1.012
WorldAfford [6]	DINO+CLIP+SAM+GPT-4	1.201	0.406	1.255	1.393	0.380	1.225	-	-	-
AffordanceLLM [38]	LLAVA-7B	-	-	-	1.463	0.377	1.070	-	-	-
Rai et al. [40]	DINO+CLIP+GPT-3.5T	1.194	0.400	1.223	1.407	0.362	1.170	-	-	-
INTRA [17]	DINOv2+ALBEF+GPT-4	1.199	0.407	1.239	1.365	0.375	1.209	-	-	-
Ours	DINO+CLIP	1.124	0.433	1.280	1.243	0.405	1.368	1.358	0.378	1.231

egocentric image, ensuring that the weakest response among the available exocentric images is considered. This is because exocentric images do not necessarily capture objects at a small scale; rather, some may be framed to emphasize only the specific regions involved in the interaction. Consequently, ρ is exploited to binarize the pixels of $A_{\text{obj}}^{\text{ego}}$, distinguishing affordance-relevant parts from other regions.

Selective Pixel Contrastive Learning. Part supervision within the egocentric view may not always be available in cases where exocentric images maintain a clear and unobstructed focus on target objects. For such circumstances, object-level learning is conducted instead to distinguish target object regions against background pixels. Thus, we utilize the hyperparameter γ_2 (equal to γ_1) to distinguish target object regions in the egocentric object affinity map $A_{\text{obj}}^{\text{ego}}$, finding that a single shared value suffices for effective separation.

Consequently, given \mathbb{J} as the index set that contains indices in which the corresponding egocentric image contains pixels in object affinity map over ρ , positive and negative sets are organized as below:

$$Q_b^+ = \begin{cases} \{\hat{F}_{b,h,w}^{\text{ego}} | A_{\text{obj},b,h,w}^{\text{ego}} > \rho_b\} & \text{if } b \in \mathbb{J}, \\ \{\hat{F}_{b,h,w}^{\text{ego}} | A_{\text{obj},b,h,w}^{\text{ego}} > \gamma_2\} & \text{otherwise,} \end{cases} \quad (7)$$

$$Q_b^- = \begin{cases} \{\hat{F}_{b,h,w}^{\text{ego}} | A_{\text{obj},b,h,w}^{\text{ego}} \leq \rho_b\} & \text{if } b \in \mathbb{J}, \\ \{\hat{F}_{b,h,w}^{\text{ego}} | A_{\text{obj},b,h,w}^{\text{ego}} \leq \gamma_2\} & \text{otherwise.} \end{cases}$$

We note that the pixels in a positive set Q_b^+ are used as anchors for pixel contrastive learning. Then, pixel contrastive learning is formulated as follows:

$$\mathcal{L}_b^{\text{pix}} = \frac{-1}{|Q_b^+|^2} \sum_{z \in Q_b^+} \sum_{p \in Q_b^+} \log \frac{\exp(z \circ p / \tau)}{\sum_{n \in (Q_b^+ \cup Q_b^-)} \exp(z \circ n / \tau)}. \quad (8)$$

This encourages the model’s attention to align with the discovered pixel-level clues, ensuring its attention precisely

corresponds to affordance-relevant regions.

3.5. Calibrating the Class Activation Map

During inference, we follow previous works [23, 28, 47] to directly employ CAM as an output localization map, representing affordable regions. However, CAM predictions often produce a Gaussian-like distribution around each salient pixel that extends beyond the actual object boundary. This occurs because convolution-based projection layers are utilized to encode local contexts, which spreads activations across pixels within the receptive fields. To this end, we apply a calibration process by performing a Hadamard product between the binarized object affinity map A and the CAM prediction to limit activations to only the salient parts. Note that the process of binarization of A is identical to the process of distinguishing target object regions in Eq. 7.

Overall Objective. Our objective involving classification loss, part-level prototypical contrastive loss, and pixel contrastive loss is expressed as $\mathcal{L} = \mathcal{L}^{\text{ce}} + \lambda_1 \mathcal{L}^{\text{proto}} + \lambda_2 \mathcal{L}^{\text{pix}}$.

4. Experiments

Evaluation Settings. For evaluation, we use two datasets, *i.e.*, AGD20K [28], and HICO-IIF [47]. Results are evaluated with Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS). These metrics evaluate the similarity and the correspondence between the distributions of prediction and ground-truth heatmaps. Also, we employ DINO ViT-S/16 and CLIP ViT-B/16 for all experiments, and set E (the number of exocentric images per egocentric image) to 3, following previous works [6, 23, 40, 47]. For hyperparameters, our loss coefficients (λ_1 and λ_2) are both set to 1. Also, for simplicity, threshold parameters (α and γ) are each set to 0.6, while the bias β and temperature τ are set to 1 and 0.5. These hyperparameters are set the same across all datasets. Further discussions on datasets and implementation details are in the Appendix.

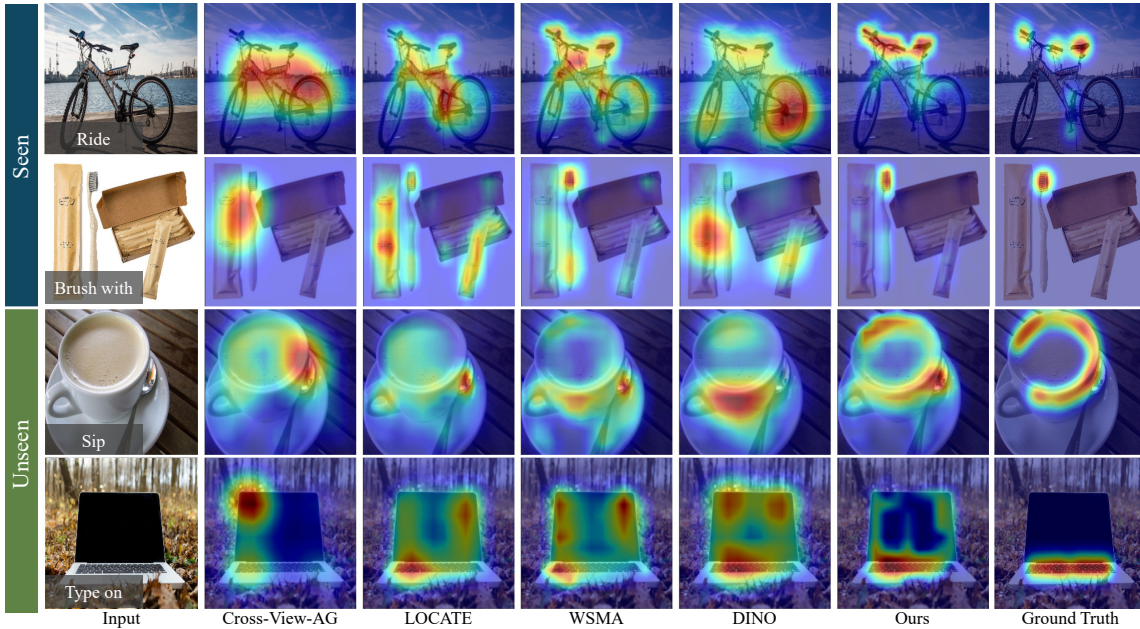


Figure 6. Qualitative comparison results of our approach and other methods in seen and unseen domains.

4.1. Comparison with the State-of-the-arts

In Tab. 1, we compare our proposed method against WSOL and WSAG methods utilizing various backbones [1, 5, 15, 19, 26, 27, 39, 45]. Models trained for object recognition, such as CLIP and WSOL methods, typically struggle with part-level grounding, as they are not optimized for identifying fine-grained affordance regions within objects. Yet, accurately locating affordance parts is as challenging for WSAG-tailored methods. Thus, methods leveraging recent VLM and LLM have emerged [6, 17, 38, 40]. Particularly, these methods often utilize LLMs to enumerate the characteristics of affordable parts within objects to improve localization with fine-grained specifications. In this work, we follow the experimental settings of [6, 23, 40, 47] to achieve a notable performance improvement across various scenarios and datasets, surpassing all previous approaches.

Particularly, we highlight the significant improvement in unseen scenarios, where novel objects are introduced for interaction. This is crucial in real-world applications, where object categories cannot be predefined. We attribute these gains to the properties of contrastive learning. First, our approach explicitly redirects attention away from background context and toward affordable parts/objects by enforcing contrastive objectives. This is especially beneficial when handling unseen objects, where the model is more prone to background distractions. Additionally, incorporating an auxiliary self-supervised objective has been shown to enhance generalizability to novel objects [3, 32], further strengthening the model’s robustness in diverse affordance scenarios.

Fig. 6 shows qualitative results in seen and unseen domains. Previous works tend to identify class-wise distinguishable parts rather than focusing on affordable regions.

Table 2. Study on model components. From left to right, we examine the benefits of object- and part-level prototypical contrastive learning (Proto.), object- and part-level pixel contrastive learning (Pixel.), and the calibration process with an object affinity map. Cali. indicates the calibration process of the localization map. Obj. and P. denote object-level and part-level learning.

	Proto.		Pixel.		Cali.	AGD20K-Seen		
	Obj.	P.	Obj.	P.		KLD	SIM	NSS
(a)	-	-	-	-	-	1.349	0.365	1.138
(b)	-	-	-	-	✓	1.271	0.394	1.162
(c)	✓	-	-	-	-	1.271	0.392	1.153
(d)	✓	-	✓	-	-	1.219	0.402	1.215
(e)	✓	-	✓	-	✓	1.198	0.419	1.198
(f)	✓	✓	-	-	-	1.164	0.416	1.290
(g)	✓	✓	✓	-	-	1.157	0.414	1.277
(h)	✓	✓	✓	✓	-	1.142	0.415	1.303
(i)	✓	✓	✓	✓	✓	1.124	0.433	1.280

For example, the bicycle frames or wheels are often highlighted instead of affordable parts for action “ride” (i.e., seat or handlebars). Our approach improves affordance precision by encouraging the model to focus on affordance-relevant parts/objects while suppressing its activation on background.

4.2. Ablation Study

Our study on component ablation is reported in Tab. 2 with fixed random seed. For our baseline, we employ the model trained solely with the classification loss. We then progressively integrate each learning strategy, noting that each component of our approach contributes positively to affordance grounding. Rows (c) and (d) demonstrate the impact of object-level learning on (a), leading to significant performance improvements in cases. These results validate our strategy to introduce object-level learning for WSAG as

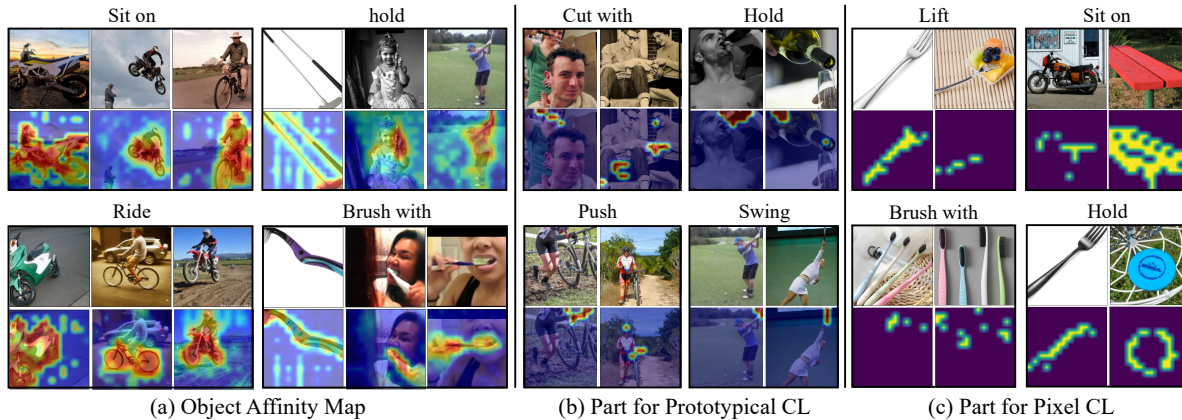


Figure 7. Visualization of discovered objects and parts used to guide the training. (a) Object affinity map A_{obj} . The leftmost sample for each class is an egocentric image and the rest are the exocentric images. (b) Affordable parts of exocentric images used for prototypical contrastive learning. (c) Affordable parts Q^+ of egocentric images used for pixel contrastive learning.

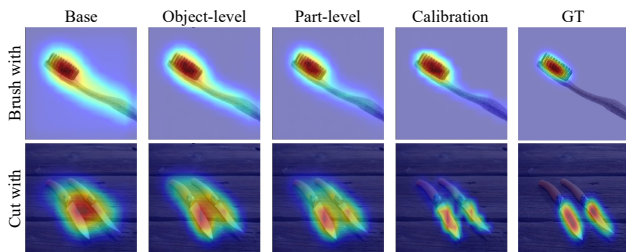


Figure 8. Analysis of the impact of each training level, *i.e.*, object and part, with qualitative results.

object-level learning mitigates the model’s confirmation bias toward unaffordable but visibly distinct parts. Additionally, the results in (f) and (h) underscore the impact of direct part-level learning, as part-level contrastive learning distinguishes affordance-relevant parts and enhances the understanding of partial details. Finally, the calibration process with an object affinity map further improves the performances by bringing two advantages: it refines the boundary and masks out the object-irrelevant activations.

The impact of each training level is further illustrated in Fig. 8. Our baseline model tends to focus on distinct parts for each class, with activations predominantly occurring on components such as the brush and center points of knives. Next, we examine the effect of object-level learning, where the model spreads activations from unaffordable parts to the general object. When part-level learning is introduced, the activation becomes more focused on regions that are more likely to be interacted with. Finally, the calibration process filters out the noisy activations surrounding the salient part regions, enhancing the accuracy of affordance grounding. These results demonstrate that the objectives of each training level are appropriately reflected.

4.3. Study on Part and Object Level Supervision

In Fig. 7, we analyze each level of training guidance to scrutinize the benefits of our object- and part-level learning. First,

(a) displays the object affinity maps. Despite that object affinity maps may have imprecise pixel-wise activation and only identify the foreground at a coarse granularity, we observe the accuracy in encompassing the action-associated objects. In (b) and (c), we visualize the detected affordable parts for exocentric and egocentric images, respectively. Specifically, (b) visualizes the detected parts within the exocentric view. Although occasional noise is present, the identified parts generally offer reliable guidance for affordance learning. In (c), we exhibit the identified affordable pixels Q^+ for part-level pixel contrastive learning on egocentric images where activated pixels generally exhibit contextual consistency. These findings affirm that our training guidance satisfiably reflects our aim of gathering reliable supervision.

5. Conclusion

To enhance part-level learning, existing approaches have employed distillation strategies to guide classifiers toward affordance-relevant parts. Yet, since affordance cues are not always distinguishable, training is often dominated by classification, which can lead the model to focus on details frequently appearing in specific classes that may not correspond to affordable parts. To address this issue, we introduced selective prototypical and pixel contrastive objectives that adaptively distinguish affordance-relevant cues from affordance-irrelevant regions at both the part and object levels. Also, we introduced a part discovery algorithm to extract affordance-relevant parts within egocentric images while incorporating a modified version of an existing approach to identify parts in exocentric images. Lastly, we applied a localized map calibration process using an object affinity map to mitigate the activation spread caused by the receptive fields in our convolution-based CAM predictions. Experimental results validate the effectiveness of our approach.

Acknowledgements

This work was supported in part by MSIT/IITP (No. RS-2022-II220680, 2020-0-01821, RS-2019-II190421, RS-2024-00459618, RS-2024-00360227, RS-2024-00437633), MSIT/NRF (No. RS-2024-00357729), KNPA/KIPoT (No. RS-2025-25393280), and SEMES-SKKU collaboration funded by SEMES.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 7
- [2] 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
- [3] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 7
- [4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 4, 7
- [6] Changmao Chen, Yuren Cong, and Zhen Kan. Worldafford: Affordance grounding based on natural language instructions. *arXiv preprint arXiv:2405.12461*, 2024. 2, 6, 7
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3
- [8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3
- [9] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In *Findings of NAACL*, 2022. 2, 5
- [10] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2219–2228, 2019. 3
- [11] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2
- [12] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2886–2895, 2021. 6
- [13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 3
- [14] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 2
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3
- [17] Ji Ha Jang, Hoigi Seo, and Se Young Chun. Intra: Interaction relationship-aware weakly supervised affordance grounding. In *European Conference on Computer Vision*, pages 18–34. Springer, 2025. 1, 2, 3, 6, 7
- [18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 7
- [20] 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. 2
- [21] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3524–3533, 2017. 3
- [22] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 3, 6
- [23] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023. 1, 2, 4, 6, 7
- [24] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024. 2

- [25] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 3
- [26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 7
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 7
- [28] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 1, 2, 6
- [29] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969, 2024. 6
- [30] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8766–8775, 2020. 3, 6
- [31] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 3
- [32] WonJun Moon, Ji-Hwan Kim, and Jae-Pil Heo. Tailoring self-supervision for supervised learning. In *European Conference on Computer Vision*, pages 346–364. Springer, 2022. 3, 7
- [33] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015. 2
- [34] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1, 6
- [35] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. 2
- [36] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11642–11651, 2021. 6
- [37] Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023. 1
- [38] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024. 1, 2, 6, 7
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7
- [40] Arushi Rai, Kyle Buettner, and Adriana Kovashka. Strategies to leverage foundational model knowledge in object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1723, 2024. 2, 6, 7
- [41] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19540–19549, 2023. 3
- [42] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Progressive proxy anchor propagation for unsupervised semantic segmentation. In *European Conference on Computer Vision*, pages 472–490. Springer, 2024.
- [43] Qing Su, Anton Netchaev, Hai Li, and Shihao Ji. Flsl: Feature-level self-supervised learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [44] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. Contextrast: Contextual contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3732–3742, 2024. 3
- [45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 7
- [46] Jun Wei, Sheng Wang, S Kevin Zhou, Shuguang Cui, and Zhen Li. Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency. In *European Conference on Computer Vision*, pages 195–210. Springer, 2022. 3
- [47] Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6324–6332, 2024. 1, 2, 6, 7
- [48] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6589–6598, 2019. 3
- [49] Shunsuke Yasuki and Masato Taki. Cam back again: Large kernel cnns from a weakly supervised object localization perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 341–351, 2024. 3
- [50] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable

- features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3
- [51] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018. 3
- [52] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10623–10633, 2021. 3
- [53] Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, and Fang Wan. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2023. 3
- [54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2