INTRA: Interaction Relationship-aware Weakly Supervised Affordance Grounding

Ji Ha Jang[1*], Hoigi Seo[1*], and Se Young Chun[1,2†]

[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{jeeit17, seohoiki3215, sychun}@snu.ac.kr

Abstract. Affordance denotes the potential interactions inherent in objects. Although recent advances in weakly supervised affordance grounding yielded promising results, there remain challenges including the requirement for paired exocentric and egocentric image dataset, and the complexity in grounding diverse affordances for a single object. To address them, we propose INTeraction Relationship-aware weakly supervised Affordance grounding (INTRA). Unlike prior arts, INTRA recasts this problem as representation learning to identify unique features of interactions through contrastive learning with exocentric images only, eliminating the need for paired datasets. Moreover, we leverage vision-language model embeddings for performing affordance grounding flexibly with any text, designing text-conditioned affordance map generation to reflect interaction relationship for contrastive learning and enhancing robustness with our text synonym augmentation. Our method outperformed prior arts on diverse datasets.

1 Motivation

Affordance refers to the perceived possible interactions based on an object's properties (*e.g.*, the rim of a wine glass affords sipping while stem of it affords holding). Affordance plays an essential role across numerous applications involving intelligent agents such as task planning, robot grasping, manipulation, scene understanding and action prediction, enabling them to provide flexible and timely responses in complex, dynamic environments.

Affordance grounding is the task to teach intelligent systems how to locate possible action regions in objects for a certain interaction. While fully supervised learning is the most straightforward approach, its reliance on costly annotations may limit its applicability across diverse contexts. Another approach is weakly supervised learning that does not require GT, but *weak* labels. In this setting, *exocentric* images illustrating human-object interactions, along with corresponding *egocentric* images depicting the objects, are provided during training. During inference, intelligent systems perform affordance grounding on the egocentric images, identifying object parts relevant to the given interactions.

* Authors contributed equally. † Corresponding author.

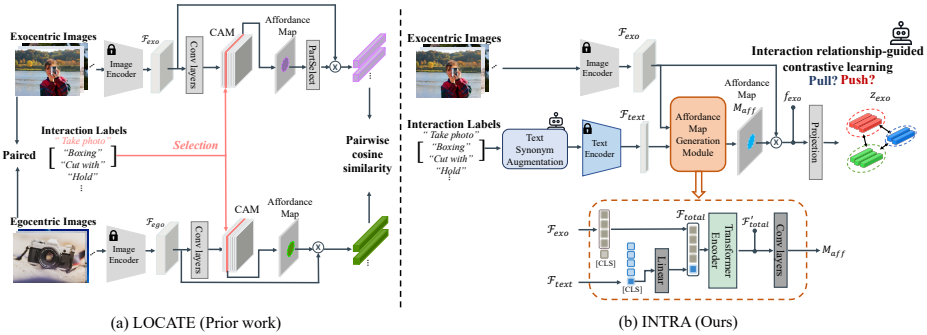

Fig. 1: Overall frameworks of (a) LOCATE and (b) INTRA (Ours).

Recent advances in weakly supervised affordance grounding (Li CVPR23, Li arXiv23, Luo CVPR22, Nagarajan ICCV19), using *pairs* of exocentric and egocentric images, have yielded great performance, but it remains a challenging task. Firstly, the requirement for current weak labels with *pairs* of exocentric and egocentric images is still strong. Secondly, a complex relationship between interactions exists, which has not been adequately addressed in prior works. Many instances in object-interaction relationships exhibit intricate many-to-many associations, occasionally with one entailing another. For example, some distinct interactions represent the same affordance regions (*e.g.*, 'wash tooth brush', 'brush with'), and there are closely related interactions that always come together (*e.g.*, 'sip' usually includes 'hold'). This complexity poses challenges in extracting interaction-relevant features based on image-level affordance labels, introducing biases towards objects in affordance grounding.

2 Method

We propose INTRA (INTeraction Relationship-aware weakly supervised Affordance grounding), a novel weakly supervised affordance grounding method to address these challenges as illustrated in Fig. 1(b). Prior arts in weakly supervised affordance grounding typically align object features of paired exocentric and egocentric images to learn interaction-related features. For example, as depicted in Fig. 1(a), LOCATE generates affordance maps from exocentric and egocentric images for a pre-determined interaction, extracts egocentric feature as well as exocentric object parts feature selected by PartSelect module, and trains the model by optimizing cosine similarity to align two object parts features. In contrast, our INTRA framework recasts the weak supervision problem as representation learning, which allows us to use *weaker* labels (*i.e.*, exocentric images only) for training so that the requirement to use *pairs* of egocentric and exocentric images is now alleviated.

Specifically, training procedure of INTRA follows these steps : First, in affordance map generation module, INTRA generates affordance map by conditioning

exocentric image features obtained using DINOv2 (Oquab arXiv24) with text feature of the interaction labels from the ALBEF text encoder (Li NeurIPS21) to utilize the semantic meanings in interaction labels and enable flexible inference on novel verbs. To enhance the robustness of text conditioning, we generate synonyms for each interaction label using LLM and randomly select the text conditioning embedding, called text synonym augmentation. Then, the exocentric image features f_{exo} corresponding to the affordance map are extracted, projected and normalized to obtain the z_{exo} using an MLP layer for contrastive learning.

The total loss for INTRA consists of two parts : 1) Interaction relationship-guided contrastive loss, using LLM to determine if interactions are positive or negative pairs based on whether they act on the same part of the object, addressing the inadequacy of treating all other interaction classes as negative pairs in affordance grounding. 2) Object-variance mitigation loss, a supervised contrastive loss on the object class, accounts for variations in affordance regions by object and context, such as holding a scissor and a cup.

3 Experimental Results

We conducted a comprehensive evaluation of INTRA, encompassing both quantitative and qualitative assessments. In Tab. 1, a comparative analysis of INTRA's performance is presented against established baselines—KLD, SIM, and NSS—on the AGD20K dataset. Notably, INTRA consistently surpasses these baselines despite trained exclusively on exocentric images.

Table 1: Quantitative results of INTRA and other works on the AGD20K dataset show that INTRA outperformed all baselines with exocentric-images only training in all metrics, KLD, SIM and NSS. ↑/↓ indicates that a higher/lower metric is better.

Prior works			Seen			Unseen		
			mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Weakly Supervised Affordance Grounding	Exo+Ego	Hotspots (Nagarajan ICCV19)	1.773	0.278	0.615	1.994	0.237	0.557
		Cross-view-AG (Luo CVPR22)	1.538	0.334	0.927	1.787	0.285	0.829
		Cross-view-AG+ (Luo arXiv22)	1.489	0.342	0.981	1.765	0.279	0.882
		LOCATE (Li CVPR22)	1.226	0.401	1.177	1.405	0.372	1.157
	Exo	**INTRA (Ours)**	1.199	0.407	1.239	1.365	0.375	1.209