Weakly Supervised Multimodal Affordance Grounding for Egocentric Images

Lingjing Xu[1], Yang Gao[1*], Wenfeng Song[2*], Aimin Hao[1]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[2]Computer School, Beijing Information Science and Technology University, China
{xulingjing, gaoyangvr, ham}@buaa.edu.cn, songwenfenga@163.com

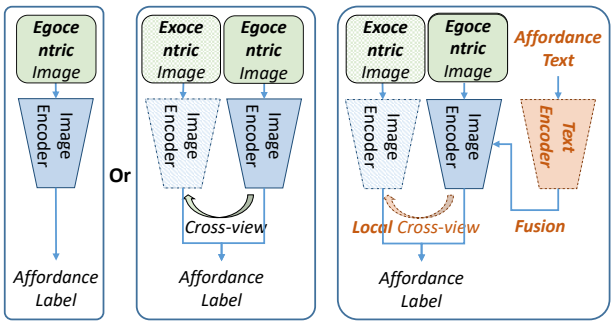
Abstract

To enhance the interaction between intelligent systems and the environment, locating the affordance regions of objects is crucial. These regions correspond to specific areas that provide distinct functionalities. Humans often acquire the ability to identify these regions through action demonstrations and verbal instructions. In this paper, we present a novel multimodal framework that extracts affordance knowledge from exocentric images, which depict human-object interactions, as well as from accompanying textual descriptions that describe the performed actions. The extracted knowledge is then transferred to egocentric images. To achieve this goal, we propose the HOI-Transfer Module, which utilizes local perception to disentangle individual actions within exocentric images. This module effectively captures localized features and correlations between actions, leading to valuable affordance knowledge. Additionally, we introduce the Pixel-Text Fusion Module, which fuses affordance knowledge by identifying regions in egocentric images that bear resemblances to the textual features defining affordances. We employ a Weakly Supervised Multimodal Affordance (WSMA) learning approach, utilizing image-level labels for training. Through extensive experiments, we demonstrate the superiority of our proposed method in terms of evaluation metrics and visual results when compared to existing affordance grounding models. Furthermore, ablation experiments confirm the effectiveness of our approach. Code:https://github.com/xulingjing88/WSMA.

Introduction

The notion of affordance, originally proposed by Gibson (Gibson 2014), posits that objects possess "action possibilities". For instance, a knife can be employed for cutting objects, while a cup can be used for drinking. However, merely knowing the purpose of an object is insufficient to enable intelligent agents to actively engage with their environment. Precise understanding of interaction locations is crucial. For example, a knife's blade is for cutting, while its handle is for gripping. This concept has garnered significant attention in the domains of robotics and computer vision, finding applications in tasks such as robotic grasping and scene comprehension.

*Corresponding author: Yang Gao and Wenfeng Song



(a) Previous Methods (b) Ours (WSMA)

Figure 1: Comparison between Previous Methods and Our Proposed Method. (a) One focused on individual object learning, and the other utilized cross-view transfer from exocentric images. (b) Our WSMA leverages local cross-view transfer from exocentric images and additionally learns from corresponding textual descriptions.

Navigating affordance knowledge is riddled with challenges. Firstly, many dominant techniques, such as those referenced in (Myers et al. 2015; Nguyen et al. 2017; Chuang et al. 2018; Do, Nguyen, and Reid 2018; Fang et al. 2018), are deeply anchored to detailed pixel-level annotations. This is particularly demanding given the complex nature of affordance regions. For example, annotating the "drink with" action for a cup entails meticulous attention to the rim area—a task that is both intricate and error-prone. Consequently, the rigors of such annotation often undermine data quality. Secondly, in real-world settings, interactions with objects are informed by both actions and language—an aspect often underserved by current methodologies. For example, Figure 1(a) indicates that traditional methods (Grabner, Gall, and Van Gool 2011; Hermans, Rehg, and Bobick 2011; Myers et al. 2015) largely view affordances as unchanging object traits, leading to segmentations predominantly based on visual appearance. Such an approach downplays the fluidity of human-object dynamics. Conversely, as depicted on the right side of Figure 1(a), newer research (Nagarajan, Feichtenhofer, and Grauman 2019; Li et al. 2023; Luo et al. 2022b) leans heavily on exocentric images of human interactions, potentially sidelining other rich data sources.

To tackle the challenges, we introduce a Weakly Supervised Multimodal Affordance Grounding approach, as illustrated in Figure 1(b). Our initial strategy mitigates the high costs associated with manual annotation by adopting a weakly supervised learning approach, where image-level labels guide the process instead of intricate pixel-level annotations. Furthermore, to harness diverse learning sources, we present an innovative methodology that leverages both exocentric images capturing human interactions and textual data for affordance knowledge acquisition. Notably, individuals learning interactions through a combination of visual and textual cues are expected to outperform those solely relying on images.

The fundamental concept driving our proposed methodology unfolds as follows: during the training phase, we employ pairs of exocentric images and corresponding affordance text to amass affordance knowledge, which is subsequently transposed to egocentric images. As depicted in Figure 1(b), our network architecture consists of three branches: Exocentric, Egocentric, and Text. The Exocentric branch is trained with exocentric images to glean affordance insights from these visual inputs. This acquired visual knowledge is then transmitted to the Egocentric branch through the HOI-Transfer Module. This integration incorporates a localization strategy, enabling the effective differentiation of various actions. In parallel, the Text branch is trained with affordance text and transfers its learned knowledge to the Egocentric branch. By amalgamating insights from both exocentric images and text, our approach achieves refined localization of affordance regions. During the testing phase, we retain the Egocentric and Text branches, employ class activation mapping (CAM) (Zhou et al. 2016) to infer the pertinent affordance regions, and subsequently enhance segmentation results using the CAM Refined Module.

In summary, our salient contributions are as follows:

- We propose a multimodal weakly-supervised framework (WSMA) designed to localize affordance regions by integrating affordance knowledge from both exocentric images and their corresponding textual descriptions, effectively transferring the acquired knowledge to egocentric images.
- We introduce the HOI-Transfer Module to extract local affordance knowledge from exocentric person-object interaction features and supervise the training of egocentric images. Concurrently, the Pixel-Text Fusion Module is incorporated to facilitate the transfer of knowledge from text to images by integrating text and egocentric image features.
- In our evaluation, we conduct experiments on two datasets: ADE20K and HICO-IIF. The experimental results clearly underscore the superiority of our proposed approach over existing methods, showcasing its optimal performance in localizing affordance regions.

Related Work

In our work, we employ a weakly supervised multimodal approach for Affordance Grounding. In this section, we survey works in domains related to our method.

Visual Affordance Grounding

Visual affordance grounding aims to locate object regions responsible for specific functionalities, thereby enhancing the comprehension of human interactions. While recent research has delved into this task, initial works relied on pixel-level annotations for supervised training (Koppula, Gupta, and Saxena 2013; Myers et al. 2015; Chuang et al. 2018; Do, Nguyen, and Reid 2018), but were limited by the complexity of annotations. The field has also yielded various weakly supervised approaches. For example, Sawatzky et al. (Sawatzky and Gall 2017) proposed a method using a minimal number of points as weak supervision, while Nagarajan et al. (Nagarajan, Feichtenhofer, and Grauman 2019) leveraged videos for learning. Recently, weak supervision through image-level labels (Luo et al. 2022b; Li et al. 2023) has emerged, primarily focusing on exocentric images for affordance learning. In contrast, our work introduces a comprehensive framework that not only attends to exocentric images but also incorporates affordance knowledge from textual sources.

Cross-view Knowledge Distillation

Knowledge distillation is a training technique in deep learning that involves the transfer of model knowledge from a teacher model to a student model (Mirzadeh et al. 2020; Chen et al. 2020). Conversely, cross-view knowledge distillation focuses on transferring knowledge across different perspectives. Research in this field has expanded in recent years (Fang et al. 2018; Sigurdsson et al. 2018; Nagarajan, Feichtenhofer, and Grauman 2019; Li et al. 2021; Luo et al. 2022b; Li et al. 2023), with some methods leveraging videos for knowledge transfer. For instance, Ego-exo (Li et al. 2021) proposes a method that uses third-person videos to uncover latent signals and predict specific attributes in egocentric views. Other methods utilize images from different perspectives for knowledge transfer. Both Cross-view-AG (Luo et al. 2022b) and LOCATE (Li et al. 2023) learn affordance knowledge from exocentric images and transfer this knowledge to egocentric images. In this paper, we employ cross-view distillation using newly designed local losses between exocentric and egocentric images.

Vision-language Models

Visual-language models aim to achieve mutual comprehension and interaction between images and natural language, establishing a close nexus between visual and textual information. An increasing number of works are dedicated to investigating this domain, with CLIP (Radford et al. 2021) being one of the most prominent examples. CLIP undertakes training on extensive image-text datasets and attains impressive performance benchmarks. Additionally, this area of research has generated numerous other significant contributions (Xu et al. 2019; Wang, Chan, and Loy 2023; Guo et al. 2023). For instance, certain investigations (Gao et al. 2021a; Zhang et al. 2021; Zhou et al. 2022) have advanced CLIP's training strategies, while others (Rao et al. 2022) focus on segmentation tasks. Inspired by these developments, we integrate CLIP's text encoder into our framework to extract textual features in our study.

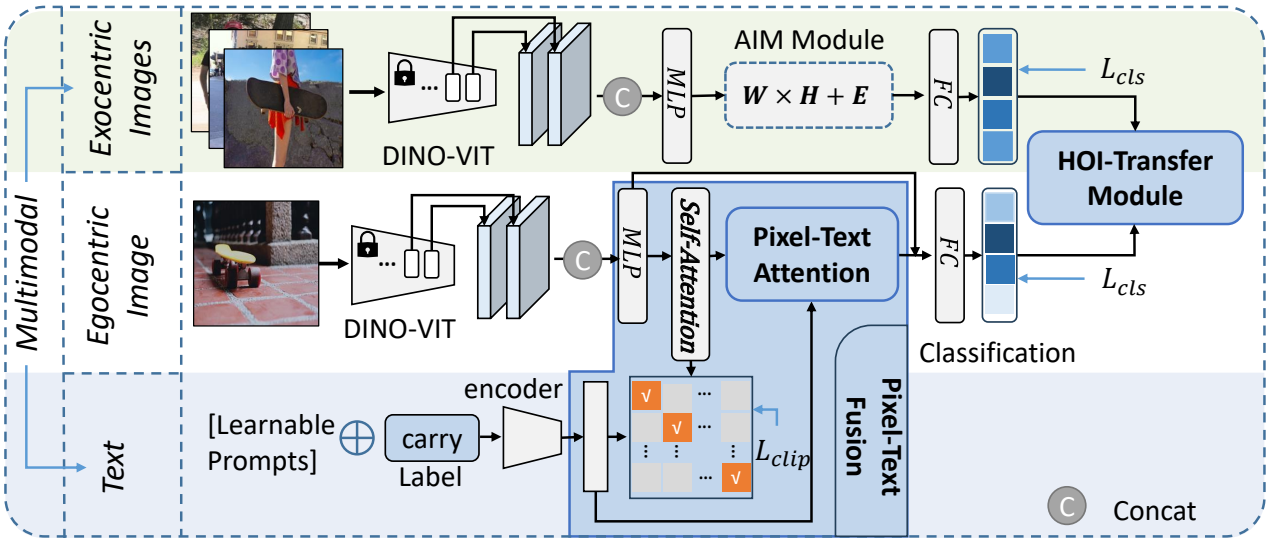


Figure 2: Overview of the Proposed Framework (WSMA). During the training stage, the proposed framework is divided into three branches: Exocentric, Egocentric, and Text. (1) The Exocentric branch extracts affordance knowledge from exocentric images and transfers it to the Egocentric branch using the HOI-Transfer Module. (2) The Egocentric branch extracts features from egocentric images and incorporates the affordance knowledge provided by the other branches. (3) The Text branch extracts features from the affordance text and fuses them into the Egocentric branch using the Pixel-Text Fusion Module.

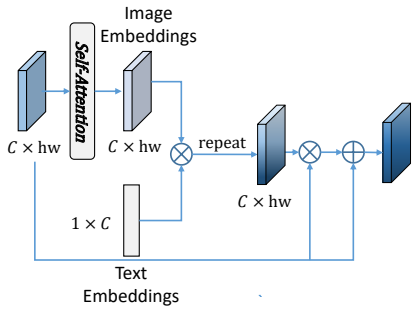


Figure 3: Detailed Description of Pixel-Text Attention. This module is designed to transfer textual knowledge.

Method

Figure 2 illustrates our Weakly Supervised Multimodal Affordance (WSMA) framework, which is designed to localize affordance regions in egocentric images. The following sections delve deeper into the intricacies of this framework.

Multimodal Fusion

While people typically learn interaction skills through demonstrations and linguistic cues, current research often overlooks this. In contrast, our approach authentically derives affordance insights from both exocentric images and descriptive text. Notably, compared to other weakly supervised methods relying on image-level labels, our approach utilizes the same labels and treats class names as textual input, without introducing new annotations or increasing the labeling workload in practical scenarios. Accordingly, our model is built upon three foundational pillars: Exocentric, Egocentric, and Text branches. The input parameters to our model include n exocentric images I_i (where i spans from 1 to n), an egocentric image I_g, and an affordance label C.

Egocentric Branch and Text Branch For a single egocentric image I_g input, we utilize the DINO-VIT (Caron et al. 2021) model for feature extraction. DINO-VIT, a self-supervised vision transformer, provides feature information pertinent to image semantic segmentation. As shown in Figure 2, our framework employs a DINO-VIT model M with $b\,(b = 12)$ blocks, yielding $\left(f_g^1, \ldots, f_g^b\right) = M\left(I_g\right)$. To enhance results comprehensively, we extract features from both the penultimate and final layers, yielding the deep feature $f_g = MLP\left(Concat\left(f_g^{b-1}, f_g^b\right)\right)$, where MLP comprises two linear layers.

With regards to the input of affordance text T (textual descriptions of affordance label C), we recognize the challenges in manual prompt design. Drawing inspiration from CoOp (Zhou et al. 2022), we introduce $m\,(m = 16)$ trainable prompts preceding T. This results in $T^{'} = [V_1] \cdots [V_m]\, T$, where $\{V_1, \ldots, V_m\}$ represent the m trainable prompts. Using the text encoder M_T from CLIP (Radford et al. 2021), we then derive text embeddings $f_t = M_T\left(T^{'}\right)$.

Pixel-Text Fusion Module To effectively merge the affordance knowledge from textual information into the Egocentric Image branch, we introduce the Pixel-Text Fusion Module. Firstly, to ensure alignment of the image features f_g and text features f_t within the same feature space, we use the following equation:

$$f_g^{'} = AttentionPool(Concat(Average(f_g), f_g)). \quad (1)$$

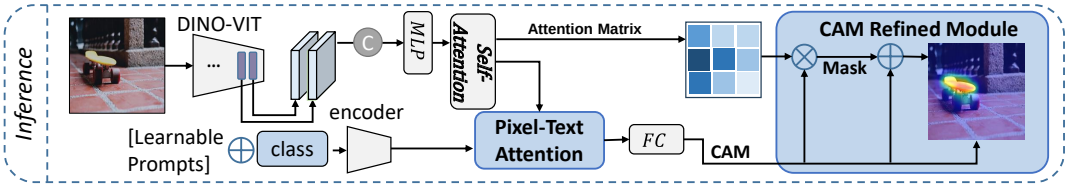


Figure 4: Description of the Testing Stage. We retain only the Egocentric and Text branches, utilizing the attention matrix from the self-attention structure within the network to optimize the results.

$AttentionPool$ is computed using a multi-head attention mechanism. Based on this, we propose the equation:

$$Z_{clip} = f_g^{'}(0) f_t^T, \tag{2}$$

where $f_g^{'}(i)$ refers to the portion of $f_g^{'}$ with channel index i. Finally, we utilize Z_{clip} to compute the cross-entropy loss L_{clip}, which guides the training.

As illustrated in Figure 3, to seamlessly merge the aligned textual features with the image features, we introduced the Pixel-Text Attention Module. This computation employs the egocentric image feature $f_g^{'} \in \mathbb{R}^{(1+hw)\times c}$ and the previously obtained text embedding $f_t \in \mathbb{R}^{1\times c}$. We start with:

$$f_{att} = f_t[f_g^{'}(1:)]^T, \tag{3}$$

where f_{att} serves as a similarity matrix bridging images and text. It functions as a subsequent attention matrix for egocentric images, directing the model's focus towards regions that resonate with the affordance text. f_{att} undergoes repetition to yield $f_{att}^{'} \in \mathbb{R}^{c\times hw}$. The final equation becomes:

$$F_g = f_g \times f_{att}^{'} + f_g, \tag{4}$$

where F_g represents the culmination of affordance knowledge transfer from the text to the egocentric images. For classification purposes, F_g is passed through a 3×3 convolutional layer followed by a fully connected layer, resulting in the classification scores c_{ego}. These scores are then used to determine the cross-entropy loss L_{cls} for optimization.

Exocentric Branch To begin with, feature extraction is conducted for the input of n exocentric images $I_i (i = \{1, \ldots, n\})$, mirroring the approach used in the Egocentric branch. Similar to the Egocentric branch, we utilize DINO-VIT for feature extraction. Subsequently, we merge the outputs from the last two layers of the network (f_i^{b-1}, f_i^b) to yield the comprehensive features $f_x^i = MLP(Concat(f_i^{b-1}, f_i^b))$.

Building upon the Affordance Invariance Mining Module (AIM) introduced in a prior study (Luo et al. 2022b), we express the comprehensive features f_x^i as $W_x \times H_x^i + E_x^i$. Here, W_x denotes the sub-feature related to human interactions, while H_x^i and E_x^i represent the coefficient matrix and individual variations of the i-th image, respectively. By minimizing E_x^i and iteratively updating W_x and H_x^i using non-negative matrix factorization (Lee and Seung 2000), we derive the shared features $F_x^i = f_x^i + Conv(W_x \times H_x^i)$ from exocentric images. Subsequently, the input features F_x^i pass through a 3×3 convolution and a fully connected layer to produce the classification scores c_{exo}. These scores, c_{exo}, are also used to compute the cross-entropy loss L_{cls}.

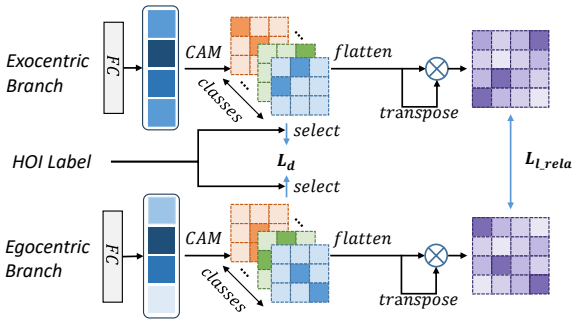


Figure 5: Detailed Description of the HOI-Transfer Module. This module is designed to transfer affordance knowledge from the Exocentric branch to the Egocentric branch.

Weakly Supervised Affordance Grounding

Given the challenges associated with annotating precise affordance regions, we rely solely on image-level annotations to calculate the loss functions. Our approach incorporates four specific losses: L_{cls}, L_{clip}, L_d, and L_{l_rela}. We have already discussed L_{cls} and L_{clip} in previous sections. Here we will delve into the details of L_d and L_{l_rela}, which are integrated within the HOI-Transfer Module.

HOI-Transfer Module Within the HOI-Transfer Module (Figure 5), we have formulated two losses, denoted as L_d and L_{l_rela}, to transfer the affordance knowledge acquired from the Exocentric branch to the Egocentric branch. Initially, we average the n exocentric features $F_x^i (i = \{1, \ldots, n\})$ to obtain F_x. We have already acquired feature F_g from the Egocentric branch and the Text branch. Inspired by Class Activation Mapping (CAM), we have devised a **local knowledge transfer** mechanism that enables better differentiation between distinct behaviors. We use CAM to calculate the weighted sum of the feature maps F_g^j, F_x^j (j represents the j-th channel) from the last convolutional layer, resulting in the affordance region heatmaps $Y_g^{C_k}, Y_x^{C_k}$ (C_k represents the k-th class) for each affordance class.

$$Y_{branch=\{g,x\}}^{C_k} = \sum_j w_j^{C_k} F_{branch=\{g,x\}}^j. \tag{5}$$

Here, $w_j^{C_k}$ represents the weights corresponding to the feature map. Subsequently, we utilize the obtained heatmaps to

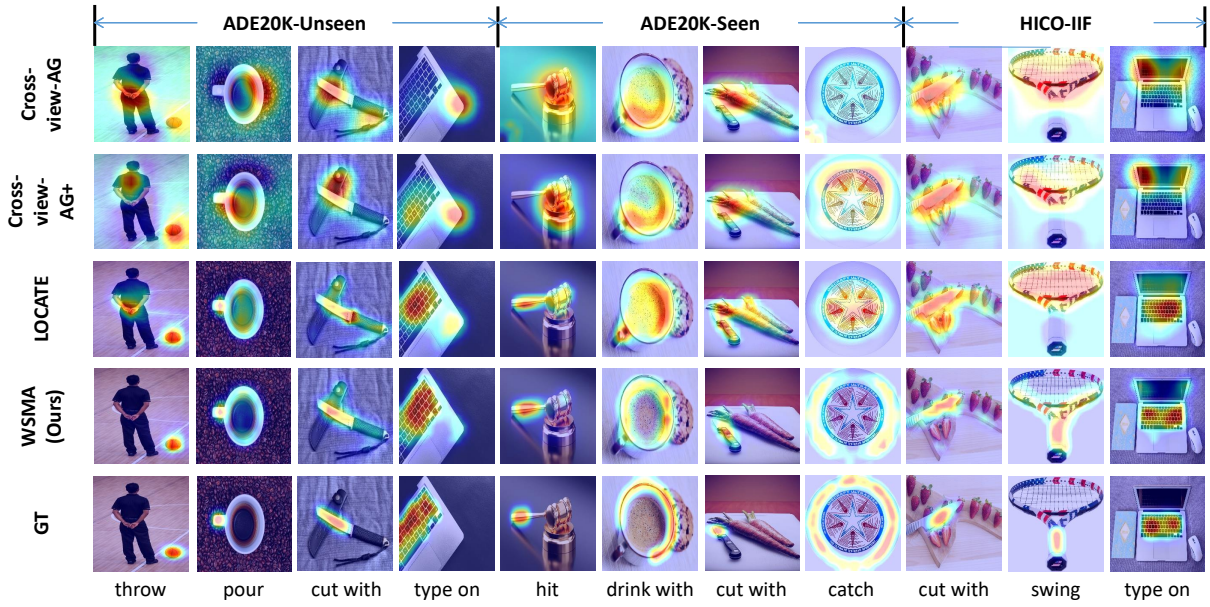


Figure 6: Qualitative Comparison with the State-of-the-art Models (Cross-view-AG(Luo et al. 2022b), Cross-view-AG+ (Luo et al. 2022a), LOCATE (Li et al. 2023)).

Method	Pub.	ADE20K-Unseen			ADE20K-Seen			HICO-IIF		
		KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
Weakly Supervised Object Localization										
SPA	CVPR21	7.425	0.169	0.262	5.528	0.221	0.357	—	—	—
EIL	CVPR20	2.167	0.277	0.330	1.931	0.285	0.522	—	—	—
TS-CAM	ICCV21	2.104	0.201	0.151	1.842	0.260	0.336	—	—	—
Affordance Grounding										
Hotspots	ICCV19	1.994	0.237	0.577	1.773	0.278	0.615	—	—	—
Cross-view-AG	CVPR22	1.787	0.285	0.829	1.538	0.334	0.927	1.779	0.263	0.946
Cross-view-AG+	—	1.765	0.279	0.882	1.489	0.342	0.981	1.836	0.256	0.883
LOCATE	CVPR23	1.405	0.372	1.157	1.226	0.401	1.177	1.593	0.327	0.966
WSMA(Ours)	**This Work**	**1.335**	**0.382**	**1.220**	**1.176**	**0.416**	**1.247**	**1.465**	**0.358**	**1.012**

Table 1: Comparisons with Other State-of-the-art Models (SPA (Pan et al. 2021), EIL (Mai, Yang, and Luo 2020), TS-CAM (Gao et al. 2021b), Hotspots (Nagarajan, Feichtenhofer, and Grauman 2019), Cross-view-AG (Luo et al. 2022b), Cross-view-AG+ (Luo et al. 2022a), LOCATE (Li et al. 2023)). ↑ indicates that a higher value is preferable, while ↓ indicates that a lower value is preferable. The experimental results that are bold and underlined represent the state-of-the-art performance.

calculate two losses. Firstly, L_d, aims to minimize the distance between the features learned in the Egocentric branch and the Exocentric branch. Accordingly, we select the corresponding heatmap based on the affordance class label C and calculate L_d as follows:

$$L_d = \left|\left|Y_g^C - Y_x^C\right|\right|. \quad (6)$$

The second loss, L_{l_rela}, addresses the fact that distinct classes in the affordance domain often exhibit overlapping characteristics. For instance, when a person holds a cup to drink water, both the "hold" and "drink with" actions are relevant, illustrating what we term action correlation. Consequently, we incorporate a loss term to transfer this action correlation knowledge from the Exocentric branch to the Egocentric branch.

$$R_{ego} = flatten(Y_g) \times flatten(Y_g^T). \quad (7)$$

R_{ego} is the correlation matrix in the Egocentric branch.

$$R_{exo} = flatten(Y_x) \times flatten(Y_x^T). \quad (8)$$

R_{exo} is the correlation matrix in the Exocentric branch.

$$L_{l\_rela} = Cosine(R_{ego}, R_{exo}). \quad (9)$$

During the training phase, the overall loss L is obtained as a weighted sum of the four individual losses.

$$L = \lambda_{cls}L_{cls} + \lambda_{clip}L_{clip} + \lambda_d L_d + \lambda_{l\_rela}L_{l\_rela}. \quad (10)$$

B_{ego}	HOI-Transfer	Pixel-Text Fusion	ADE20K-Unseen			ADE20K-Seen			HICO-IIF		
			KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
✓	✗	✗	1.707	0.287	0.973	1.430	0.345	1.093	1.860	0.255	0.770
✓	✓	✗	1.471	0.330	1.180	1.297	0.375	1.178	1.785	0.286	0.650
✓	✗	✓	1.531	0.338	1.095	1.277	0.393	1.169	1.690	0.320	0.970
✓	✓	✓	**1.335**	**0.382**	**1.220**	**1.176**	**0.416**	**1.247**	**1.465**	**0.358**	**1.012**

Table 2: Ablation Experiments of Different Modules (B_{ego} is the Egocentric branch)

$\lambda_{cls}, \lambda_{clip}, \lambda_d, \lambda_{l\_rela}$ represent the weights corresponding to the respective losses.

Finally, during the inference phase (Figure 4), we retain only the Egocentric branch and the Text branch. We input a single egocentric image and associate it with the affordance label C. For instance, if you aim to determine the affordance region of an object for the action "catch", you can input the image of the object along with the corresponding textual label "catch". Through the application of CAM, we generate the corresponding heatmap H. To further enhance the heatmap, we utilize the CAM Refined Module. In this module, the network's self-attention mechanism extracts the attention matrix Q, which is then used to refine H.

$$H^{'} = Mask \times Q \times H + H, \tag{11}$$

where $H^{'}$ represents the refined heatmap, and the $Mask$ is

$$\text{Mask} = \left\{ \begin{array}{ll} 1 & \text{if } H > \text{threshold} \\ 0 & \text{else} \end{array} \right\}. \tag{12}$$

We set a $threshold$ to remove less crucial portions from $Q \times H$, thus focusing more on the important parts.

Experimental Results

Datasets and Evaluation Metrics

We use the Affordance Grounding Dataset (AGD20K) (Luo et al. 2022b), which is a comprehensive dataset containing various viewpoints, specifically, 20,061 exocentric and 3,755 egocentric images. These images represent 36 unique affordance categories. We conduct evaluations under two distinct settings: "Seen" and "Unseen". The "Seen" setting includes object categories from the training set in the test set, whereas the "Unseen" setting incorporates object categories that are not present in the training set. In addition to AGD20K, we have assembled a new dataset, HICO-IIF, by selecting specific subsets from the HICO-DET (Chao et al. 2018) and IIT-AFF (Nguyen et al. 2017) datasets. More details about these datasets are available in the Appendix.

To measure the alignment between experimental outcomes and the ground truth, we use three metrics: Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS). The Appendix provides a comprehensive overview of each metric.

Implementation Details

For the backbone of both the egocentric and exocentric branches, we use the pre-trained DINO-ViT-S, keeping its weights frozen during the training process. DINO-ViT-S is pre-trained using unsupervised learning on the ImageNet dataset (Deng et al. 2009). In the case of the exocentric branch, we simultaneously process input from three exocentric images. The text branch, on the other hand, employs the pre-trained text encoder from the CLIP model as its backbone network. We set the hyperparameters λ_{cls}, λ_{clip}, λ_d, and λ_{l_rela} to 1, 1, 0.5, and 0.5 respectively, while the $threshold$ is fixed at 0.2. Further details regarding parameter configurations can be found in the Appendix.

Quantitative and Qualitative Comparisons

We benchmark our method against three weakly supervised object localization models and four cutting-edge affordance grounding models. The results of these models are tabulated in Table 1. Notably, our proposed method, WSMA, outperforms all other compared methods. Specifically, when juxtaposed with the current leading affordance grounding model, LOCATE, WSMA demonstrates superior performance across all three metrics. These experimental results underscore the potency of our approach in transferring learned knowledge from exocentric images and text to egocentric images, consequently improving heatmap accuracy.

Furthermore, to provide a more detailed visual analysis of the differences among various model results, we conducted a qualitative comparison. As illustrated in Figure 6, it can be inferred that under the "Unseen" setting, WSMA achieves improved precision in localization by effectively mitigating environmental influences. For instance, under the "throw" label, WSMA accurately pinpoints the region of the basketball, whereas other models exhibit varying degrees of sensitivity to environmental factors introduced by individuals in the scene. Similarly, across the other two datasets, WSMA consistently yields results that closely align with the ground truth, outmatching other models in comparison.

Ablation Study

To validate the efficacy of the proposed modules, we conducted comprehensive ablation experiments, with results summarized in Table 2. Methods relying solely on the Egocentric branch exhibited the lowest performance. Analyzing the three evaluation metrics revealed that integrating the HOI-Transfer Module or the Pixel-Text Fusion Module led to varying degrees of improvement. Ultimately, the combined integration of both modules achieved the highest performance. This accomplishment can be attributed to the efficient extraction of knowledge from exocentric images and

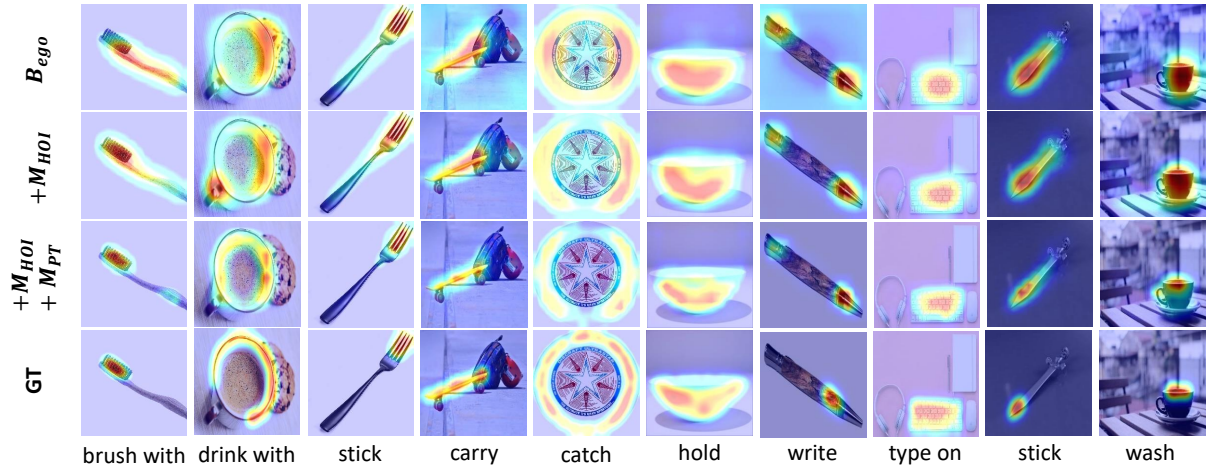


Figure 7: Ablation Experiments' Visualization. M_{HOI} is the HOI-Transfer Module. M_{PT} is the Pixel-Text Fusion Module.

L_d	L_{g_rela}	L_{l_rela}	ADE20K-Unseen			ADE20K-Seen			HICO-IIF		
			KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
✗	✗	✗	1.531	0.338	1.095	1.277	0.393	1.169	1.690	0.320	0.970
✓	✗	✗	1.512	0.340	1.103	1.248	0.398	1.216	1.594	0.328	1.022
✓	✓	✗	1.499	0.339	1.144	1.255	0.399	1.205	1.624	0.329	0.997
✓	✗	✓	**1.335**	**0.382**	**1.220**	**1.176**	**0.416**	**1.247**	**1.465**	**0.358**	**1.012**

Table 3: Ablation Experiments of Two Losses in the HOI-Transfer Module.

textual data through these two modules, followed by the seamless transfer of this knowledge to egocentric images. As a consequence, exceptional results have been attained.

Visual comparisons are presented in Figure 7. When compared to using only the Egocentric branch, the inclusion of the HOI-Transfer Module enhances the accuracy of identifying approximate affordance region locations. For example, by incorporating the HOI-Transfer Module, more precise attention can be directed towards the head of the toothbrush. Moreover, Figure 7 demonstrates the beneficial effect of the Pixel-Text Fusion Module in localizing affordance regions, effectively eliminating interference from other parts and achieving precise localization.

Furthermore, we conducted ablation experiments on two loss functions within the HOI-Transfer Module (see Table 3). We examined the impact of including or excluding L_d in the experiments. Additionally, regarding the HOI correlation loss, we compared the efficacy of two distinct loss functions, namely L_{g_rela} and L_{l_rela}, with the latter already introduced in the methodology section. L_{g_rela} was introduced in a prior work (Luo et al. 2022b), where it calculates action relevance using classification scores. Analyzing Table 3, we observe that the experimental results are superior when incorporating L_d. When L_{g_rela} is added in addition to L_d, the three evaluation metrics do not simultaneously achieve superior results. However, if we replace L_{g_rela} with L_{l_rela}, all three evaluation metrics significantly outperform the results without its inclusion. Notably, L_{l_rela} (Figure 5) first identifies the heatmaps for each category before performing correlation calculations. This suggests that the superior performance of L_{l_rela} is attributed to the heatmaps containing more informative and valuable information compared to simple classification scores.

Conclusion

This work introduces a novel weakly supervised multimodal framework, WSMA, for localizing affordance regions. The main idea is to learn affordance knowledge from both exocentric images and affordance text. The framework utilizes the HOI-Transfer Module to extract affordance knowledge from exocentric images, while the Pixel-Text Fusion Module integrates knowledge from text into egocentric images. During testing, the framework takes only the egocentric image and its corresponding affordance text to determine object affordance regions. WSMA demonstrates superior performance compared to state-of-the-art methods.

However, our work still has limitations due to the lack of complex interaction images in existing public datasets. For instance, there may be situations where an image contains multiple objects of different categories but with the same affordance label. To address these challenges, we plan to improve datasets and tackle the challenges arising from such complex interactions.

Acknowledgments

This paper is supported by the National Key R&D Program of China (2023YFC3604500), National Natural Science Foundation of China (62002010, 62102036), Beijing Natural Science Foundation (L232102, 4222024), the Beijing Science and Technology Plan Project (No. Z221100007722001, Z231100005923039), RD Program of Beijing Municipal Education Commission (KM202211232003), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2022A02, No.VRLAB2022C06).

References

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Chen, D.; Mei, J.-P.; Wang, C.; Feng, Y.; and Chen, C. 2020. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3430–3437.

Chuang, C.-Y.; Li, J.; Torralba, A.; and Fidler, S. 2018. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 975–983.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Do, T.-T.; Nguyen, A.; and Reid, I. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, 5882–5889. IEEE.

Fang, K.; Wu, T.-L.; Yang, D.; Savarese, S.; and Lim, J. J. 2018. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2139–2147.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021a. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021b. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2886–2895.

Gibson, J. J. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.

Grabner, H.; Gall, J.; and Van Gool, L. 2011. What makes a chair a chair? In *CVPR 2011*, 1529–1536. IEEE.

Guo, Z.; Zhang, R.; Qiu, L.; Ma, X.; Miao, X.; He, X.; and Cui, B. 2023. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 746–754.

Hermans, T.; Rehg, J. M.; and Bobick, A. 2011. Affordance prediction via learned object attributes. In *IEEE international conference on robotics and automation (ICRA): Workshop on semantic perception, mapping, and exploration*, 181–184. Citeseer.

Koppula, H. S.; Gupta, R.; and Saxena, A. 2013. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8): 951–970.

Lee, D.; and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Li, G.; Jampani, V.; Sun, D.; and Sevilla-Lara, L. 2023. LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10922–10931.

Li, Y.; Nagarajan, T.; Xiong, B.; and Grauman, K. 2021. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6943–6953.

Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; and Tao, D. 2022a. Grounded affordance from exocentric view. *arXiv preprint arXiv:2208.13196*.

Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; and Tao, D. 2022b. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2252–2261.

Mai, J.; Yang, M.; and Luo, W. 2020. 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*, 8766–8775.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.

Myers, A.; Teo, C. L.; Fermüller, C.; and Aloimonos, Y. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1374–1381. IEEE.

Nagarajan, T.; Feichtenhofer, C.; and Grauman, K. 2019. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8688–8697.

Nguyen, A.; Kanoulas, D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. 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)*, 5908–5915. IEEE.

Pan, X.; Gao, Y.; Lin, Z.; Tang, F.; Dong, W.; Yuan, H.; Huang, F.; and Xu, C. 2021. Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11642–11651.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Sawatzky, J.; and Gall, J. 2017. Adaptive binarization for weakly supervised affordance segmentation. In *Proceedings of the IEEE international conference on computer vision workshops*, 1383–1391.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9062–9069.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.