Learning Affordance Grounding from Exocentric Images (Supplementary Material)

Hongchen Luo[1,‡*] Wei Zhai[1,‡] Jing Zhang[2†] Yang Cao[1,4,†] Dacheng Tao[3,2]
[1] University of Science and Technology of China
[2] The University of Sydney [3] JD Explore Academy
[4] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{lhc12, wzhai056}@mail.ustc.edu.cn, jing.zhang1@sydney.edu.au,
forrest@ustc.edu.cn, dacheng.tao@gmail.com

Contents

A. Dataset

A.1. Collection Details

Our exocentric images are mainly derived from COCO [27], and HICO [3]. Besides, we also collect some images from PAD [28], OPRA [12] and UCF101 [47]. We choose images or video frames according to verbs and affordance categories. Our goal is to capture affordance cues from diverse individual differences determined by the object's intrinsic properties, rather than imitating interactions from individual persons. Therefore, for the video dataset, we select only a portion of the video frames interacting with the object. On the other hand, we download $5,867$ high-quality free images ($2,112$ exocentric images and $3,755$ egocentric images) from this link, which we retrieve according to the affordance category as well as the object category, and manually choose the images that satisfy the requirements. More examples of exocentric and egocentric images are shown in Fig. 1.

A.2. Annotation Details

For the test set part-level labels, we refer to the OPRA dataset [12] for the annotation of interaction regions and the annotation routine from previous visual saliency works [1, 2, 22]. By observing the interactions between humans and objects in the exocentric images, we label the egocentric images with points of different densities according to the probability of interaction between the human and object regions. In generating the mask, we apply a Gaussian blur to each labeled point and normalize it to obtain the affordance heatmaps. More examples of annotations for the affordance region in egocentric images are shown in Fig. 1.

A.3. Dataset Division

In the "Seen" setting, all exocentric images are used as the training set, while for egocentric images, we use $3,022$ as training images and 733 as test images. In the "Unseen" setting, we first select the affordance category containing several object classes. Twenty-five affordance categories satisfy the requirements, covering 47 object categories. We choose 35 classes as the training set and 12 classes as the

*This work was done during an internship at JD Explore Academy.
†Corresponding author. ‡ Equal contributions.

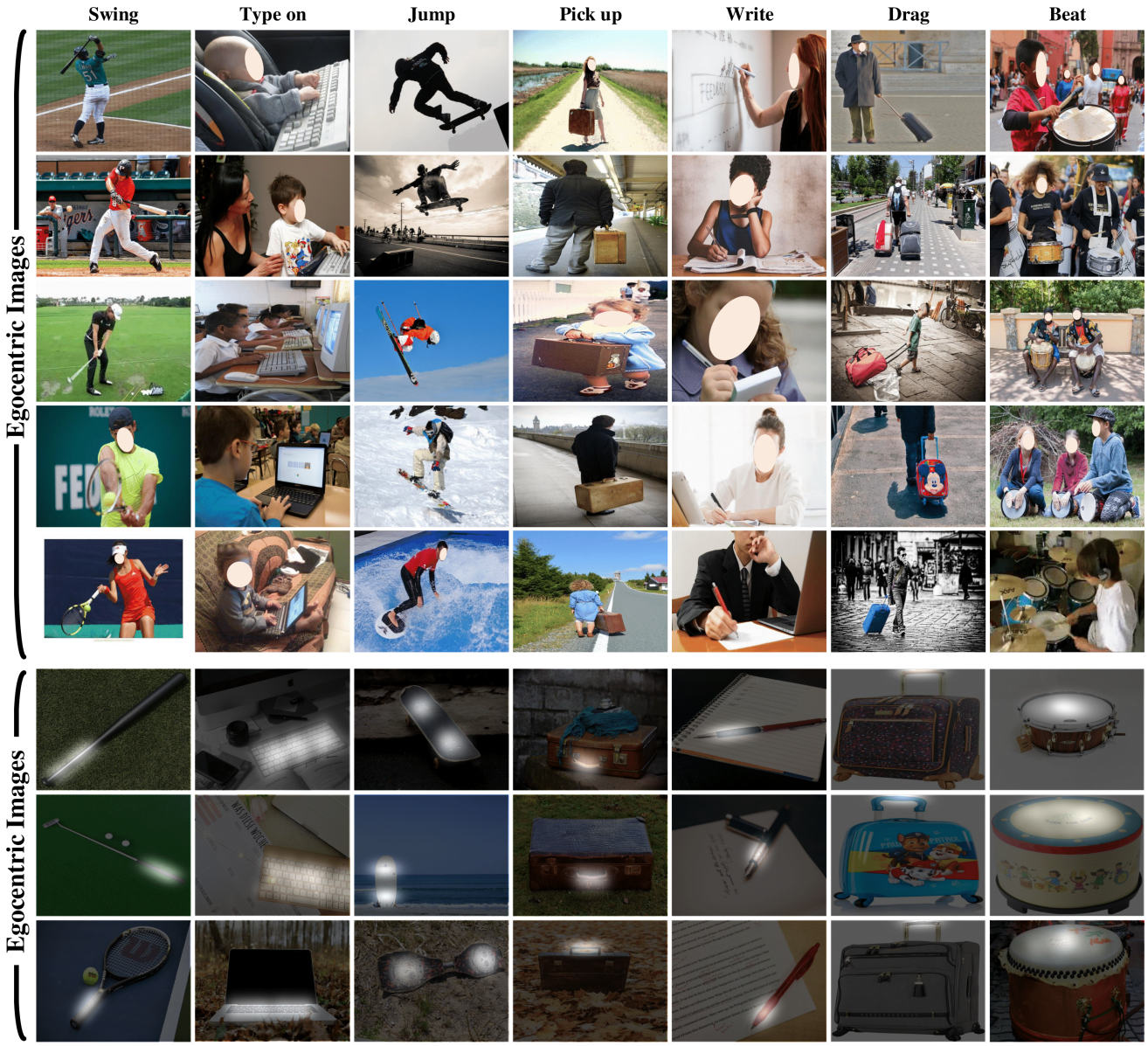

Figure 1. **Dataset examples.** More examples of exocentric and egocentric images in AGD20K.

test set, as shown in Table 1.

Besides, to explore the model's predictive performance for different scales of affordance regions, we split the test set into "Big" subsets if the proportion of the mask to the whole image is greater than 0.1, "Middle" subsets if the ratio is between 0.03 and 0.1, and "Small" subsets for the remaining part. Most of the test images fall in the "Middle" subset.

A.4. More Statistical Analysis

Table 2 shows comparisons of the AGD20K dataset with other affordance-related datasets. Our AGD20K dataset is unique in that it explicitly takes into account the exocentric to egocentric viewpoint transformation and considers higher quality images, richer affordance classes and more complex scenarios. We count the number of images of object categories contained under each affordance category in the exocentric images (as shown in Fig. 2). It can be seen from this that the distribution between affordance categories is unbalanced and satisfies the characteristics of a long-tailed distribution. In future work, we will consider further optimizing the model [9, 21, 26] to obtain better prediction results for affordance regions based on the data distribution characteristics. Our definition of affordance regions for each of

Table 1. **Dataset Division.** Under the "Unseen" setting, the training and test sets are divided by object category.

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

the 36 categories is shown in Table 3, along with the object categories contained under each affordance category. It indicates that the affordance co-relation is independent of the semantic class of objects and is commonly present in objects. Specifically, there are co-relations between "Hold" and most affordances, such as cutting something, swinging a racket, or pouring a glass of water, all of which require the presence of a "Hold" position to complete these interactions.

B. Implementation Details

B.1. Metrics

Previous works mainly segmented precise affordance regions [**?**, 7, 32, 37], while our task considers a weakly supervised setting that predicts the affordance heatmap using only the affordance category labels. Referring to the works of Demo2Vec [12], Hotspots [33], and Mlnet [8], we adopt heatmaps give a better description of the "action possibilities" ($i.e.$, affordance) and use KLD, SIM, and NSS to evaluate the probability distribution correlation between the predicted affordance heatmap and GT.

– **KLD** [2]: Kullback-Leibler Divergence (KLD) is used to measure the distribution difference between the prediction map and the target map. Given a prediction map P and a ground truth map Q^D, $KLD(\cdot)$ is computed as follows:

$$KLD(P, Q^D) = \sum_i Q_i^D log(\epsilon + \frac{Q_i^D}{\epsilon + P_i}), \quad (1)$$

where ϵ is a regularization constant.

– **SIM** [48]: The similarity metric (SIM) measures the similarity between the prediction map and the ground truth map. Given a prediction map P and a continuous ground truth map Q^D, $SIM(\cdot)$ is computed as the sum of the minimum values at each pixel, after normalizing the input maps:

$$\begin{aligned} &SIM(P, Q^D) = \sum_i min(P_i, Q_i^D), \\ &where \quad \sum_i P_i = \sum_i Q_i^D = 1. \end{aligned} \quad (2)$$

– **NSS** [41]: The Normalized Scanpath Saliency measures the correspondence between the prediction map and the ground truth, and it treats false positives and false negatives symmetrically. Given a prediction map P and a binary ground truth map Q^D, $NSS(\cdot)$ computes the average normalized prediction at ground truth locations:

$$\begin{aligned} &NSS(P, Q^D) = \frac{1}{N} \sum_i \hat{P} \times Q_i^D, \\ &where \quad N = \sum_i Q_i^D \quad and \quad \hat{P} = \frac{P - \mu(P)}{\sigma(P)}. \end{aligned} \quad (3)$$

$\mu(P)$ and $\sigma(P)$ represent the mean and standard deviation of P, respectively.

B.2. Comparison Methods

– Mlnet [8]: Unlike previous works that predict saliency maps directly from the last layer of the convolution neural network, the model fuses feature extracted from different CNN layers. Their method contains three main blocks: feature extraction CNN, feature encoding network (weighting of low and high features), and a prior learning network. The method achieves promising results in all datasets for saliency detection.

– DeepGazeII [24]: Unlike other saliency models, DeepGazeII does not perform additional fine-tuning of the VGG features and only trains some output layers to predict saliency on top of VGG.

– EgoGaze [20]: It is a hybrid gaze prediction model that exploits both the visual saliency of bottom-up and task-dependent attention transition and is the first work

Table 2. **Statistics of related datasets and the proposed AGD20K dataset.** Part: part-level annotation. HQ: high-quality annotation. BG: the background is fixed or from general scenarios. Exo&Ego: whether to transfer from exocentric to egocentric view. ♯Obj: number of object classes. ♯Aff: number of affordance classes. ♯Img: number of images.

Dataset	Pub	Year	HQ	Part	BG	Exo & Ego?	♯Obj.	♯Aff.	♯Img.
UMD [32]	ICRA	2015	✗	✓	Fixed	✗	17	7	30,000
[44]	CVPR	2017	✗	✓	Fixed	✗	17	7	3,090
IIT-AFF [37]	IROS	2017	✗	✓	General	✗	10	9	8,835
ADE-Aff [7]	CVPR	2018	✓	✓	General	✗	150	7	10,000
PAD [28]	IJCAI	2021	✓	✗	General	✗	72	31	4,002
AGD20k (Ours)	CVPR	2022	✓	✓	General	✓	50	36	23,816

to explore the attention transition model in the egocentric gaze prediction task and achieves state-of-the-art results in gaze prediction.

For the three models for saliency detection described above, we use models trained on the saliency-related datasets for testing in the same way as [33].

– EIL [29]: Since CAM-based approaches for weakly supervised object localization (WSOL) highlight only the most discriminative regions, not the whole object. Therefore, this thesis proposes a novel adversarial erasing technique jointly exploring highly response class-specific areas and less discriminative regions to obtain a more complete object region.

– SPA [38]: It explores how to extract object structure information during the training phase of the classification network and proposes a structure-preserving activation (SPA) method that leverages the structure information incorporated in the convolutional features for WSOL.

– TS-CAM [14]: It proposes a token semantic coupled attention map (TS-CAM) approach that captures long-range visual dependencies using a self-attention mechanism. By introducing a semantic decoupling module, the semantic-aware tokens and the semantic-agnostic attention map are combined to jointly leverage semantic and localization information to better localize object regions.

For the three weakly supervised object localization models mentioned above, we only utilize exocentric images for training. In the main paper, we also conduct experiments and analyse the effect of using both exocentric and egocentric images on the performance of the weakly supervised object localization models.

– Hotspots [33]: It is a weakly supervised way to learn the affordance of an object through video, and affordance grounding is achieved only through action labels.

C. Experiments

C.1. Different Classes

To investigate the performance of the models on different affordance categories, we show the results of the KLD metrics for each category in the "Seen" setting. The experimental results are shown in Table 4. **Bold** and Underline indicate the best and the second-best scores, respectively. Our model achieves the best results under most affordance categories, which demonstrates the superiority of our method in locating affordance regions. Specifically, our model obtains more accurate results for both the affordance categories "Hold" and "Cut with", where there are some co-relations, demonstrating that our approach can enhance the network's ability to perceive and locate affordance regions by aligning the co-relation matrices of the outputs of the two views. For affordance categories such as "Pour" and "Carry", where the interaction habits of different humans are quite diverse, our method still outperforms all other models, illustrating the effectiveness of the affordance invariance mining module in extracting affordance-specific features for affordance area localization. Due to the long-tailed distribution characteristic of the dataset, we calculate the means of the KLD metrics for all categories to compare different models from the category balanced perspective. Our method still outperforms other methods, which shows that our method can capture affordance-related features on most of the categories and achieve better affordance region localization performance despite the unbalanced data distribution.

C.2. Different Scales

To explore the capability of the model in localizing affordance regions on objects of different scales, we divided the test set into three subsets of "Big", "Middle" and "Small". The results are shown in Table 5. Our model outperforms the other methods in most of the three settings. Most of the objects in the "Big" subset have a large affordance region, and the differences in human-object interaction are generally quite large, making it difficult to fully locate the affor-

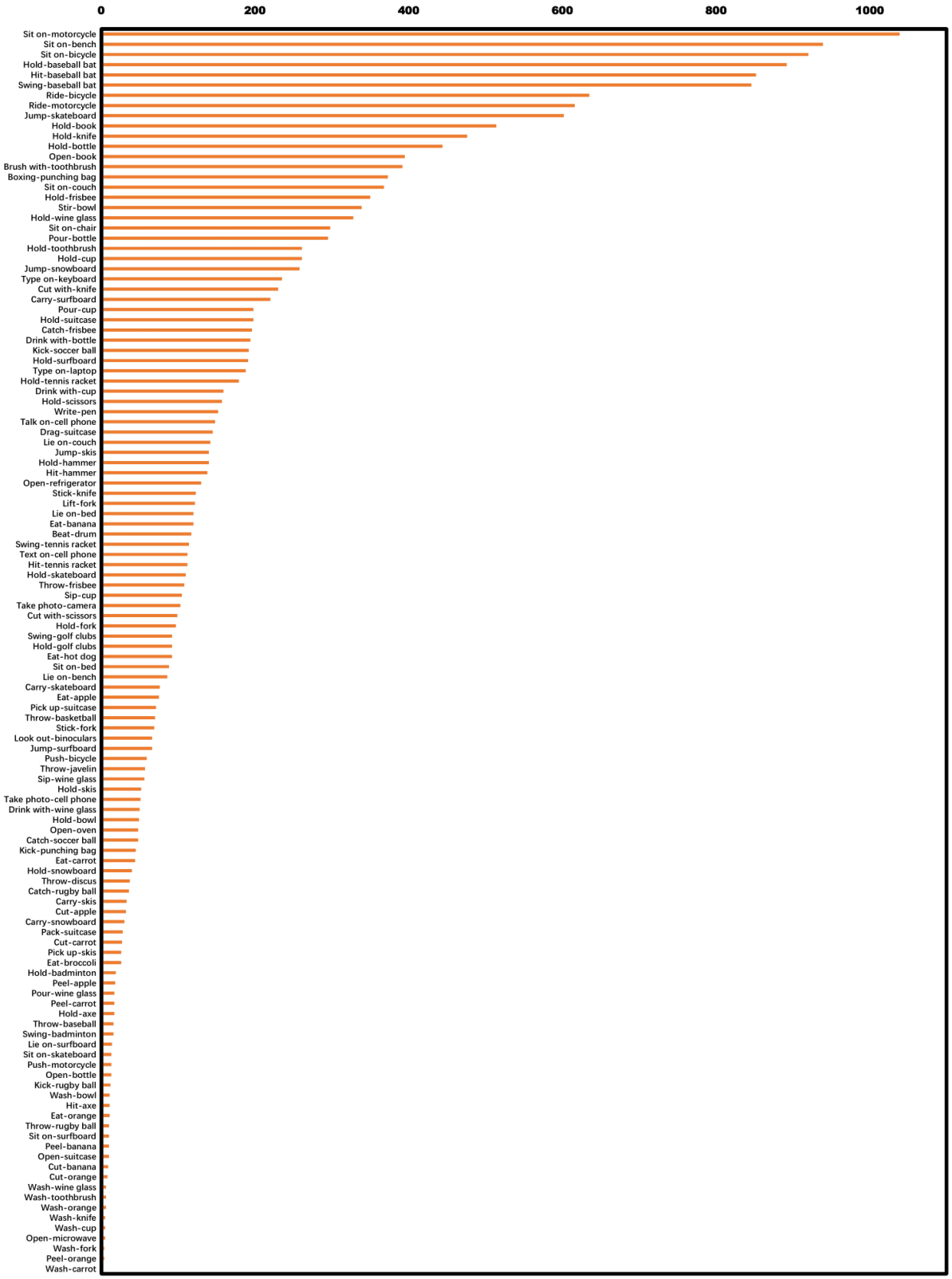

Figure 2. **Dataset statistics.** The statistics of the number of exocentric images in the object category contained under each affordance category.

Table 3. **Category definitions.** Definition of the affordance regions in the AGD20K dataset and the object categories contained in each affordance category.

Class	Description	Object Class
Beat	The object regions that can be played by beating a surface to produce a sound.	drum
Boxing	The object regions hit by boxing sports.	punching bag
Brush with	The object regions that can be used to brush the teeth.	toothbrush
Carry	Regions that interact during moving objects.	skateboard, skis, snowboard, surfboard
Catch	The interaction region that catches the object process.	frisbee, rugby ball, soccer ball
Cut	The object region that can be cut.	apple, banana, carrot, orange
Cut with	The object regions that have the ability to cut other Obj.	knife, scissors
Drag	The object regions that can be dragged.	suitcase
Drink with	The regions in which the drinking process interacts with objects.	bottle, cup, wine glass
Eat	The object regions that can be eaten.	apple, banana, broccoli, carrot, orange
Hit	Indicates object regions that can be used to strike other objects.	axe, baseball bat, hammer, tennis racket
Hold	Refers to the object regions that can be held in the hand.	axe,badminton racket, baseball bat, book, bottle, cup, fork, frisbee, golf clubs, hammer, knife, scissors, skateboard, snowboard, suitcase, bowl, surfboard, tennis racket, skis, toothbrush, wine glass
Jump	The object regions that allow rapid movement by allowing people to jump on surfaces.	skateboard, skis, snowboard, surfboard
Kick	The object regions that can be kicked in direct contact with the foot.	soccer ball, punching bag, rugby ball
Lie on	The object regions with a large surface space that allow a person to lie down.	bed, bench, couch, surfboard
Lift	The regions that interact during the lifting of the object.	fork
Look out	The object regions that can be used for seeing at a distance.	binoculars
Open	The region contacted during the opening of the object.	book, bottle, microwave, oven, refrigerator, suitcase
Pack	The object regions that can be used for packing.	suitcase
Peel	The regions in which objects are peeled.	apple, banana, carrot, orange
Pick up	The regions that can be picked up.	skis, suitcase
Pour	The regions that hands interact with during the pouring of liquids.	bottle, cup, wine glass
Push	The object regions that can be pushed forward.	bicycle, motorcycle
Ride	The regions contacted during riding.	bicycle, motorcycle
Sip	The regions that the mouth contacts with the container during the sipping process.	cup, wine glass
Sit on	The object regions that can be used to sit.	bed, bench, bicycle, chair, couch, motorcycle, skateboard, surfboard
Stick	The sharper regions that can pierce into other objects.	knife, fork
Stir	The object regions that can be used for mixing.	bowl
Swing	Object regions that a person interacts with by swinging their arm.	badminton racket, baseball bat golf clubs, tennis racket
Take photo	Interaction regions that can take pictures of people.	camera, cell phone
Talk on	The object regions contacted during calling.	cell phone
Text on	The object regions that can be used to edit text.	cell phone
Throw	The object regions that a person uses to throw.	baseball, basketball, discus, frisbee, javelin, rugby ball
Type on	The object regions that can be used for typing.	keyboard, laptop
Wash	The object regions that are cleaned with water.	bowl, carrot, cup, fork, knife, orange, toothbrush, wine glass
Write	The object regions that can be used for writing.	pen

Table 4. **Different classes.** The test results under the KLD metric for different affordance categories.

Classes	Mlnet [8]	DeepGazeII [24]	EgoGaze [20]	EIL [29]	SPA [38]	TS-CAM [14]	Hotspots [33]	Ours
Beat	4.021	**1.046**	4.388	1.326	9.548	1.547	1.440	1.231
Boxing	4.475	1.413	2.918	**1.087**	6.554	1.474	1.398	1.270
Brush with	6.215	2.385	4.935	3.003	8.043	2.642	2.154	**2.040**
Carry	6.228	1.841	3.521	1.799	5.902	1.825	1.556	**1.443**
Catch	5.356	1.448	3.519	1.019	6.641	1.011	0.917	**0.595**
Cut	2.947	0.951	3.128	0.970	5.059	1.408	1.078	**0.805**
Cut with	6.664	2.200	4.389	1.917	3.945	2.179	2.055	**1.652**
Drag	15.01	4.562	4.827	**3.877**	7.884	3.764	4.241	4.046
Drink with	4.497	2.067	4.268	2.253	7.683	2.300	1.943	**1.748**
Eat	2.778	0.944	3.052	1.039	6.400	1.373	0.959	**0.819**
Hit	8.305	2.168	4.882	2.145	6.141	2.172	1.929	**1.787**
Hold	6.762	2.071	4.671	2.008	3.006	1.628	1.770	**1.594**
Jump	5.852	1.876	3.840	2.017	8.454	2.049	1.622	**1.579**
Kick	4.353	1.169	2.758	**0.908**	3.277	1.070	1.239	0.914
Lie on	4.767	1.602	2.921	1.377	4.006	1.370	1.566	**1.039**
Lift	7.922	2.377	4.319	2.269	8.708	2.309	**2.038**	2.389
Look out	2.348	1.347	3.475	2.267	9.12	**1.216**	1.402	1.316
Open	7.108	2.172	5.416	1.984	2.892	1.867	1.916	**1.512**
Pack	3.224	**0.816**	2.684	1.272	10.85	1.145	1.486	1.002
Peel	3.715	1.021	3.373	1.032	7.594	1.494	1.147	**0.742**
Pick up	7.481	2.779	3.186	**2.608**	11.70	2.967	2.751	2.619
Pour	3.945	1.937	3.512	1.943	4.426	2.139	1.809	**1.432**
Push	4.240	2.757	5.839	3.490	13.38	**2.655**	2.904	3.000
Ride	2.364	**1.812**	3.878	2.548	5.730	2.023	2.220	1.890
Sip	3.755	1.794	3.868	1.964	7.284	2.094	1.798	**1.564**
Sit on	3.491	1.934	4.058	2.240	3.748	1.959	2.161	**1.745**
Stick	7.977	2.867	7.444	2.976	7.276	2.864	2.635	**2.334**
Stir	3.032	**0.859**	3.674	1.406	9.239	1.325	1.126	0.916
Swing	9.248	2.478	6.723	2.486	6.720	2.420	2.178	**2.161**
Take photo	2.549	1.153	3.939	1.362	7.407	1.468	1.148	**0.996**
Talk on	5.805	**1.269**	3.787	2.085	10.24	1.940	1.597	1.426
Text on	3.667	1.381	3.599	1.652	8.039	2.000	1.804	**1.356**
Throw	5.991	1.536	3.903	1.330	6.743	1.262	1.037	**0.817**
Type on	2.549	1.125	3.113	1.074	2.144	1.042	0.963	**0.503**
Wash	5.205	1.499	5.297	2.056	10.62	2.200	1.471	**1.397**
Write	4.455	2.252	4.120	2.446	6.245	2.386	1.992	**1.580**
Mean	5.231	1.803	4.095	1.923	7.045	1.905	1.763	**1.534**

dance region. In the "Small" subset, the region of interaction is generally small and challenging to locate precisely. Our method has no particular degradation in performance in either case and still accurately predicts the affordance region of the object.

C.3. Different Hyper-parameters

We investigate the influence of $\boldsymbol{T}$ on the performance of the model (as shown in Table 6. The hyper-parameter $\boldsymbol{T}$ in the ACP strategy has a smoothing effect on the category correlation distribution and plays a preservation role for affordance co-relation information. It shows that the value of $\boldsymbol{T}$ has a relatively significant impact on the performance of the model, and too large a $\boldsymbol{T}$ may cause harmful effects. We also investigate the effect of the rank r of the dictionary matrix W in the AIM module on model performance (as shown in Table 7), with different ranks representing the number of bases of the human-object interaction subfeatures. The best results are obtained when $r = 64$. A smaller r (*e.g.*, 16 or 32) may lead to poor results due to

Table 5. **Different scales.** We divide the test set into "Big", "Middle" and "Small" subsets according to the ratio of mask to the whole image, and test the performance of the model in different scales of objects.

	Scale	Big			Middle			Small		
	Method	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
Seen	Mlnet [8]	5.382	0.389	0.375	4.939	0.280	0.640	5.598	**0.176**	0.704
	DeepGazeII [24]	1.216	0.450	0.417	1.742	0.271	0.732	2.718	0.143	0.574
	EgoGaze [20]	3.400	0.339	0.235	4.174	0.226	0.376	4.941	0.124	0.328
	EIL [29]	1.047	0.461	0.389	1.794	0.284	0.710	3.057	0.123	0.231
	SPA [38]	5.745	0.317	0.222	4.990	0.228	0.440	6.076	0.118	0.297
	TS-CAM [14]	1.039	0.424	0.166	1.814	0.248	0.401	2.652	0.132	0.352
	Hotspots [33]	0.986	0.448	0.408	1.738	0.265	0.672	2.587	0.149	0.683
	Ours	**0.766**	**0.533**	**0.652**	**1.485**	**0.322**	**1.040**	**2.373**	0.175	**0.927**
Uneen	Mlnet [8]	4.441	0.426	0.491	4.554	0.281	0.657	6.058	**0.151**	0.545
	DeepGazeII [24]	0.936	0.464	0.574	1.776	0.269	0.728	2.879	0.128	0.392
	EgoGaze [20]	2.902	0.349	0.339	4.100	0.220	0.376	5.292	0.124	0.312
	EIL [29]	1.199	0.393	0.271	1.906	0.246	0.482	3.082	0.113	0.116
	SPA [38]	8.299	0.259	0.254	6.938	0.186	0.333	7.784	0.095	0.144
	TS-CAM [38]	1.238	0.351	0.072	1.970	0.208	0.236	2.766	0.113	0.124
	Hotspots [33]	1.015	0.425	0.548	1.872	0.242	0.605	2.693	0.134	0.544
	Ours	**0.884**	**0.500**	**0.728**	**1.595**	**0.303**	**0.945**	**2.558**	0.147	**0.692**

the number of bases being too small to fully represent the interactions' sub-features. And a larger r (*e.g.*, 128 or 256) may also lead to poor results, possibly due to the redundancy of information caused by the excessive number of bases. The impact of different exocentric images on model performance is shown in Table 8, with a relatively significant effect on model performance as N increases from 1 to 3. It indicates that the affordance invariance mining module can capture affordance-specific cues from multiple images, playing a prominent role in affordance region prediction.

C.4. Difference Attentions

We also explore the impact of different attention modules on the model performance, where VQ [11] and CD [10] are two different update algorithms for matrix decomposition, and Non-local [51], A^2Net [6] and Co-attention [11] are general forms of attention (as shown in 9). Our methods outperform the general attention modules, mainly due to the complexity of category demarcation in high-dimensional spaces, *i.e.*, using affordance category labels only as supervision would lead to over-parameterization of the general attention module, making it challenging to learn affordance-specific features from the complex and diverse interactions. In contrast, low-rank matrix decomposition maps feature into a set of compact dictionary bases and reconstruct features so that features in high-dimensional space are redistributed in low-dimensional subspaces, effectively solving the over-parameterization problem. In contrast to other matrix decomposition methods, we constrain W and H to be non-negative. The reconstruction results are a summation of bases only, which is more suitable for modeling the realistic sub-features of human-object interactions. Thus, we use the non-negative matrix decomposition to model human-object interaction features more explicitly and obtain affordance-specific cues.

C.5. Different Visualization Techniques

We also investigate the impact of different visualisation techniques (Grad-CAM [45], Grad-CAM++ [4], and XGrad-CAM [4]) on the prediction results (as shown in Table 10). In the "Seen" setting, the gap between CAM, Grad-CAM and Grad-CAM++ is not too significant, but in the "Unseen" setting, Grad-CAM and Grad-CAM++ have a larger gap with CAM in terms of KLD metrics. While XGrad-CAM produces results that are far from the ground truth in both the "Seen" and "Unseen" settings. It suggests that excessively complex heatmap calculations may lead to negative effects. In future work, we can improve the visualization technique to obtain more accurate prediction results according to the characteristics of the affordance region.

C.6. More Visual Results

We also show affordance heatmaps for several relevant domain best models (Hotspots [33], EIL [29] and DeepGazeII [24]) in the "Seen" and "Unseen" settings (as shown in Fig. 3 and Fig. 4). These results demonstrate the effectiveness of our method of explicitly extracting affordance-specific cues from exocentric interactions and transferring them to egocentric images, enabling more precise localization of an object's affordance region.

Table 6. **The influence of T.** We investigate the impact of the hyper-parameter T in the affordance co-relation preserving strategy on model performance.

	T=?	KLD ↓	SIM ↑	NSS ↑
Seen	0.5	1.685	**0.335**	0.837
	1 (Ours)	**1.538**	0.334	**0.927**
	2	1.623	0.328	0.845
	3	1.672	0.318	0.810
	4	1.737	0.282	0.765
Unseen	0.5	1.981	0.272	0.645
	1 (Ours)	**1.787**	**0.285**	**0.829**
	2	1.903	0.263	0.692
	3	1.969	0.252	0.603
	4	2.043	0.230	0.519

Table 7. **The influence of the rank r.** We investigate the influence of the rank r of the dictionary matrix W in the affordance invariance mining module on the performance of the model.

	Rank r	KLD ↓	SIM ↑	NSS ↑
Seen	16	1.642	0.322	0.832
	32	1.598	0.329	0.874
	64 (Ours)	**1.538**	0.334	**0.927**
	128	1.660	**0.336**	0.843
	256	1.647	0.314	0.803
Unseen	16	1.873	0.280	0.721
	32	1.827	0.277	0.782
	64 (Ours)	**1.787**	**0.285**	**0.829**
	128	1.886	0.279	0.732
	256	1.866	0.276	0.719

Table 8. **The influence of N.** We investigate the effect of different numbers of exocentric images N on the performance of the model.

	N=?	KLD ↓	SIM ↑	NSS ↑
Seen	1	1.704	0.287	0.811
	2	1.623	0.313	0.909
	3 (Ours)	**1.538**	0.334	**0.927**
	4	1.558	**0.335**	0.896
	5	1.617	0.322	0.823
Unseen	1	1.926	0.257	0.693
	2	1.871	0.277	0.782
	3 (Ours)	**1.787**	**0.285**	**0.829**
	4	1.835	0.282	0.764
	5	1.846	0.263	0.749

D. Related Works

D.1. Weakly Supervised Object Localization

Weakly supervised object localization aims to learn object localization given only category labels at the image level. Zhou et al. [59] use class activation (CAM) techniques to enable a classification-trained network to classify and localize class-related object regions simultaneously. To solve the problem of localized discriminative regions or irrelevant background regions caused by CAM-based methods, Singh et al. [46] force the network to focus on other relevant parts by randomly hiding patches of the training images. Furthermore, Mai et al. [29] introduce a novel adversarial erasing technique that explores both high response category-specific regions and low discriminative regions. Xue et al. [53] propose a divergent activation to learn complementary visual cues. Pan et al. [38] present a higher-order self-correlation to obtain structure-preserving localization maps. In a recent study, Gao et al. [14] use transformer to model long-term dependencies to avoid partial activation. Unlike the above methods, we consider learning affordance-specific clues from exocentric interactions and transferring them to egocentric images, using only the affordance labels as supervision, which is different from the general weakly supervised object localization task setting. Moreover, the affordance attribute is not equivalent to the semantic category of the object. The affordance category may contain multiple object categories, while an object may also have multiple different affordance regions. Consequently, this imposes a significant challenge in extracting affordance-related features from exocentric images due to the large variation in human interactions.

Table 9. **Different attention.** We compare the performance of using different forms of attention in our model.

	Attention Type	KLD ↓	SIM ↑	NSS ↑
Seen	VQ [16]	1.787	0.269	0.595
	CD [10]	1.905	0.245	0.377
	Non-local [51]	1.831	0.263	0.585
	A^2Net [6]	2.008	0.232	0.315
	Co-attention [11]	1.685	0.279	0.886
	NMF (Ours)	**1.538**	**0.334**	**0.927**
Unseen	VQ [16]	2.020	0.251	0.561
	CD [10]	2.017	0.259	0.575
	Non-local [51]	2.021	0.221	0.508
	A^2Net [6]	1.991	0.259	0.603
	Co-attention [11]	1.947	0.253	0.625
	NMF (Ours)	**1.787**	**0.285**	**0.829**

D.2. Knowledge Distillation

Knowledge distillation is usually used to distill knowledge from a larger and deeper network of teachers to a smaller network, mainly for model compression and acceleration. It is mainly divided into response-based knowledge, feature-based knowledge, and relationship-based knowledge [15]. Response-based knowledge [18, 30, 42,

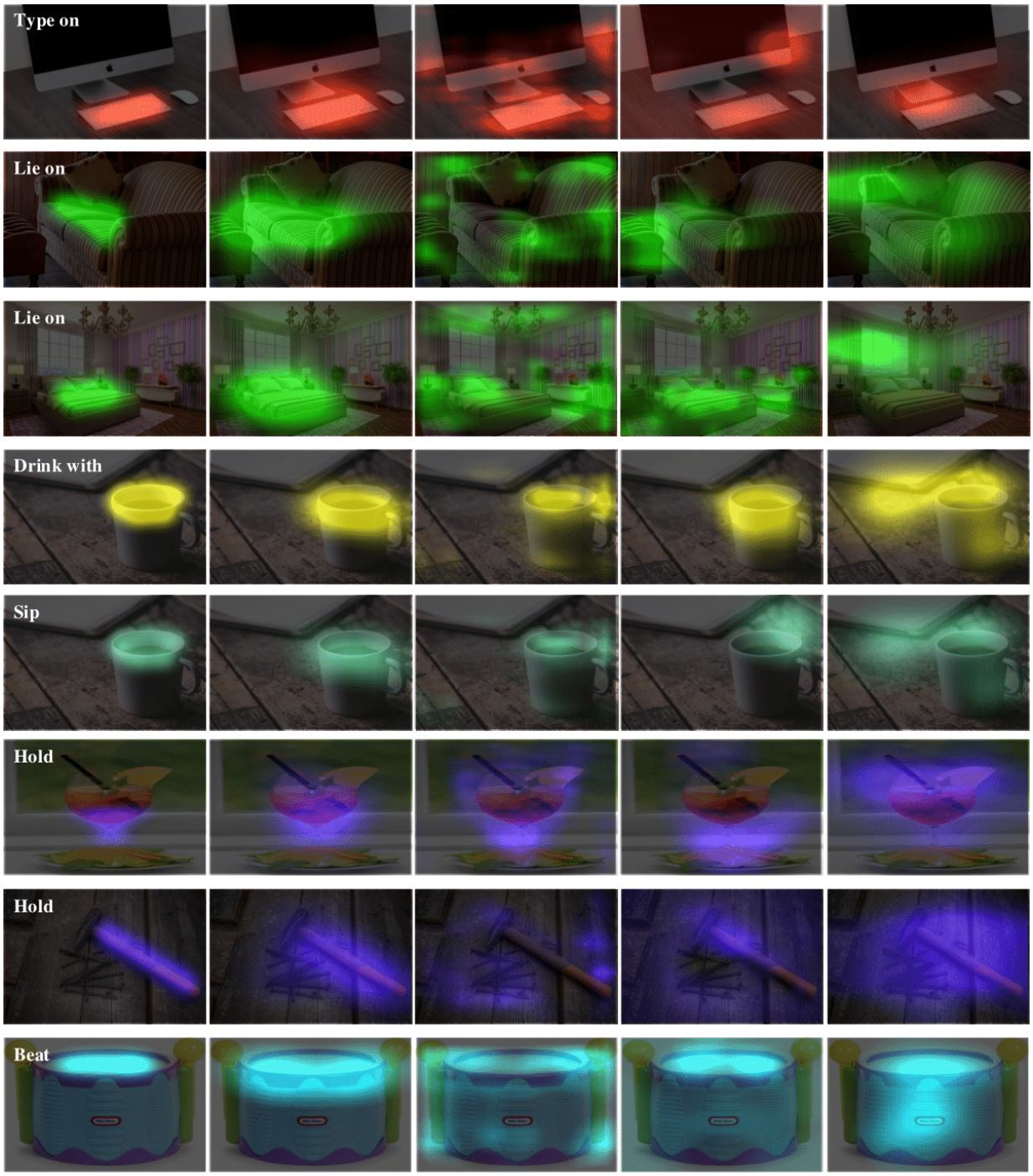

Figure 3. **Visualization results.** The affordance heatmap predictions obtained by representative models in each domain (Hotspots [33], EIL [29], DeepGazeII [24]) in the "Seen" setting.

52, 54, 58] is generally learned from the output layer of the teacher model, and the student model can obtain the informative dark knowledge contained in the teacher model from the soft labels. Response-based knowledge is generally learned from the output layer of the teacher model, and the student model can obtain the informative dark knowledge contained in the teacher model from the soft labels. Feature-based approaches [5, 17, 23, 40, 43, 50, 56] mainly consider the knowledge of the teacher model from the intermediate layers, which is an extension of response-based knowledge

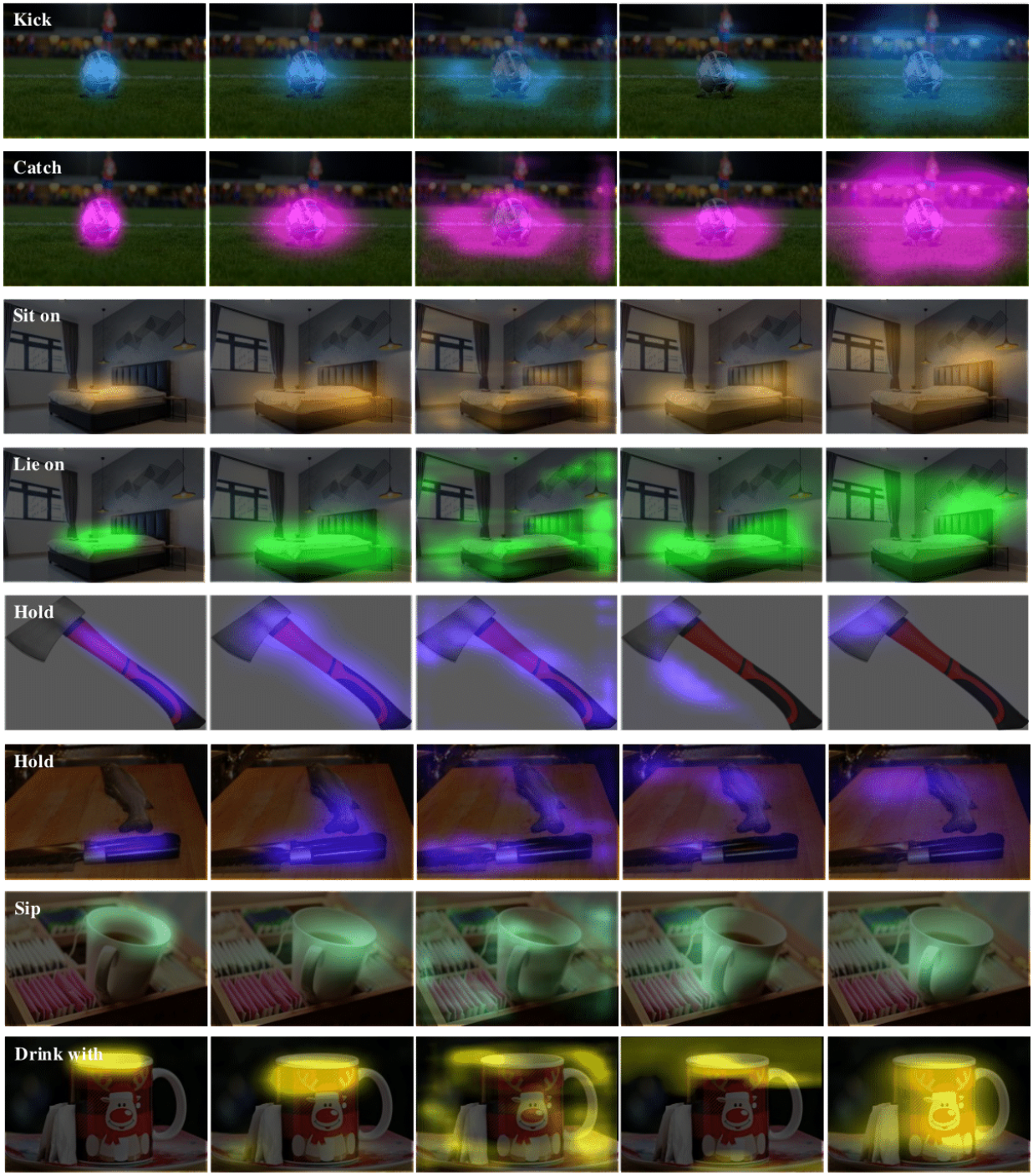

Figure 4. **Visualization results.** The affordance heatmap predictions obtained by representative models in each domain (Hotspots [33], EIL [29], DeepGazeII [24]) in the "Unseen" setting.

for thinner and deeper network training. Relation-based knowledge [25, 39, 49, 55, 57] explores the relationships between different network layers or data samples. While in this paper, we refer to the technique of knowledge distillation to transfer the affordance-specific knowledge extracted from exocentric images to egocentric images. Furthermore, we introduce the affordance co-relation preserving strategy to distill the co-relation between affordances from the exo-

Table 10. **Different visualization techniques.** We compare the performance of using different visualization techniques to generate affordance heatmaps in our model.

	Visualization	KLD ↓	SIM ↑	NSS ↑
Seen	CAM [59] (Ours)	1.538	0.334	**0.927**
	Grad-CAM [45]	1.545	0.341	0.910
	Grad-CAM++ [4]	**1.535**	**0.343**	0.915
	XGrad-CAM [13]	2.499	0.329	0.801
Unseen	CAM [59] (Ours)	**1.787**	0.285	0.829
	Grad-CAM [45]	1.930	0.296	0.829
	Grad-CAM++ [4]	1.910	**0.298**	**0.839**
	XGrad-CAM [13]	3.079	0.286	0.696

Figure 5. **Failure case.** We show the failure case of the model prediction results in the case of multiple objects or slender objects.

centric branch to the egocentric branch.

E. Limitations

Some failure cases are shown in Fig. 5. The results generated by our method may activate the intermediate background areas between multiple objects. For objects with slender structures, the background-irrelevant regions may also be activated. In future work, we consider enhancing the corresponding regions of the affordance class during training, ignoring the irrelevant background regions, and refining the generated results in line with the characteristics of affordance to obtain more accurate predictions.

F. Potential Applications

- **Action Anticipation.** By predicting the affordance of the objects that people are interacting with and the objects in the range of human activities, we can provide an efficient search space for the anticipation of possible future actions and improve the accuracy of the prediction of possible future actions [36].

- **Human-Object Interaction.** By activating different regions on the object where the interaction occurs, the model can guide an agent to focus on the key regions where human-object interaction can be inferred to improve the prediction accuracy. And it can be used in zero-shot scenarios to infer the category of human-object interactions from local region features of objects even if no examples are seen during the training process [19].

- **Self-exploration of Agents.** Interacting with the environment is a fundamental skill for embodied intelligence. When an intelligent agent arrives at a new environment, it should be able to understand the possible interactions with objects in the environment through *a priori* knowledge already acquired (trained models) and actively interact with the environment to acquire new knowledge. Such an ability has applications in robot navigation and robot grasping [31, 34, 35].

References

[1] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 1

[2] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 1, 3

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 381–389. IEEE, 2018. 1

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 8, 12

[5] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7028–7036, 2021. 10

[6] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-nets: Double attention networks. In *NeurIPS*, 2018. 8, 9

[7] 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 (CVPR)*, pages 975–983, 2018. 3, 4

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016. 3, 7, 8

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[10] Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175, 2001. 8, 9

[11] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8, 9

[12] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Lim J. Joseph. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[13] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020. 12

[14] 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 International Conference on Computer Vision (ICCV)*, pages 2886–2895, October 2021. 4, 7, 8, 9

[15] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 9

[16] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE transactions on information theory*, 44(6):2325–2383, 1998. 9

[17] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019. 10

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 9

[19] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 12

[20] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 754–769, 2018. 3, 7, 8

[21] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020. 2

[22] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 1

[23] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *arXiv preprint arXiv:1802.04977*, 2018. 10

[24] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 3, 7, 8, 10, 11

[25] Seunghyun Lee and Byung Cheol Song. Graph-based knowledge distillation by multi-head attention network. *arXiv preprint arXiv:1907.02226*, 2019. 11

[26] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1

[28] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection. *arXiv preprint arXiv:2108.03658*, 2021. 1, 4

[29] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8766–8775, 2020. 4, 7, 8, 9, 10, 11

[30] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 9

[31] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176. IEEE, 2021. 12

[32] 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. 3, 4

[33] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 8688–8697, 2019. 3, 4, 7, 8, 10, 11

[34] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. *arXiv preprint arXiv:2008.09241*, 2020. 12

[35] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 12

[36] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 12

[37] 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. 3, 4

[38] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11642–11651, 2021. 4, 7, 8, 9

[39] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 11

[40] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. *arXiv preprint arXiv:2012.14022*, 2020. 10

[41] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005. 3

[42] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1355–1364, 2019. 9

[43] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 10

[44] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8, 12

[46] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 9

[47] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[48] Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision (IJCV)*, 7(1):11–32, 1991. 3

[49] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 11

[50] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 325–342. Springer, 2020. 10

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 8, 9

[52] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 9

[53] 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. 9

[54] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Training deep neural networks in generations: A more tolerant teacher educates better students. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5628–5635, 2019. 9

[55] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 11

[56] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 10

[57] Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018. 11

[58] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 9

[59] 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 (CVPR)*, pages 2921–2929, 2016. 9, 12