Leverage Interactive Affinity for Affordance Learning
(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@mail.ustc.edu.cn, jing.zhang1@sydney.edu.au,
{wzhai056,forrest}@ustc.edu.cn, dacheng.tao@gmail.com

Contents

A. Dataset

In this section, we describe more details of the dataset collection, annotation, and the process of dividing **`Seen`** / **`Unseen`**, as well as provide more statistical analysis.

A.1. Collection Details

Since AGD20K [16] contains a large number of exocentric images of human-object interaction and egocentric images corresponding to interacting objects, it is available to provide pairs of interactive images and non-interactive images for our Contact-driven Affordance Learning (CAL) dataset. On the other hand, the PADv2 [30] dataset contains 39 affordance categories, covering a total of 103 object categories, from which we considered collecting a large number of data to enrich the CAL dataset further. Specifically, we choose 23 affordance categories frequently occurring daily. And then, we retrieve the corresponding images from the above two datasets according to the affordance categories and guarantee that the number of interactive and non-interactive images is approximately equivalent. As a result, we obtain $2,689$ interactive images and $2,569$ non-interactive images. The interactive and non-interactive image examples are shown in Fig. 1 and Fig. 2, respectively.

Although the dataset contains a small number of images, the input is determined by interactive and non-interactive image pairs. The combination of two images provides a multitude of variations (the trainset can combine a total of 170K pairs). Besides, affordance is not a particularly difficult property for humans to understand, making it easy to scale the number of images. The number of input pairs will increase exponentially with the number of images.

A.2. Annotation Details

We also assign object labels to the interactive and non-interactive images according to their object categories. Due to affordance means the "action possibilities" on objects, and it is relatively simple for a human with normal cognitive abilities to perceive. Following Gebru et al [8], we employ 10 random volunteers from the laboratory, ensuring that 3 volunteers annotated each image. We establish the following rules before labeling: (1) Only annotate the tools as the dataset focuses on labeling interactions with tools. The passive interaction of objects (such as the "cut" involves hand contact with vegetables) will be considered in future work. (2) Label the contact regions of "Hands", "Feet" etc. (3) Ignore the differences between the left & right in hands and feet. (4) The points are marked more intensively for regions where interaction is more frequent. During dataset processing, we select the average of all annotated for broadly similar regions and the majority of annotated for controversial

[*] Corresponding author. [‡] Equal contributions.

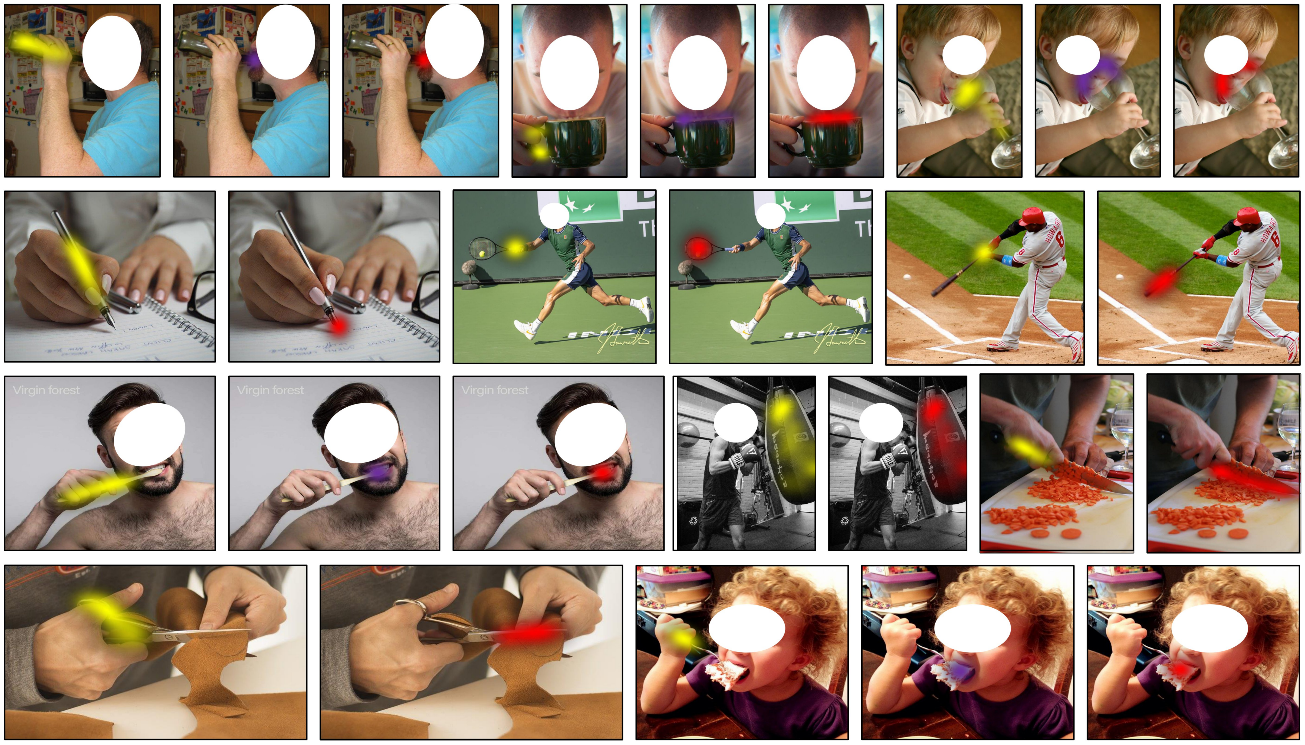


Figure 1. **Interactive image examples.** Some examples of interactive images and annotations from the CAL dataset.

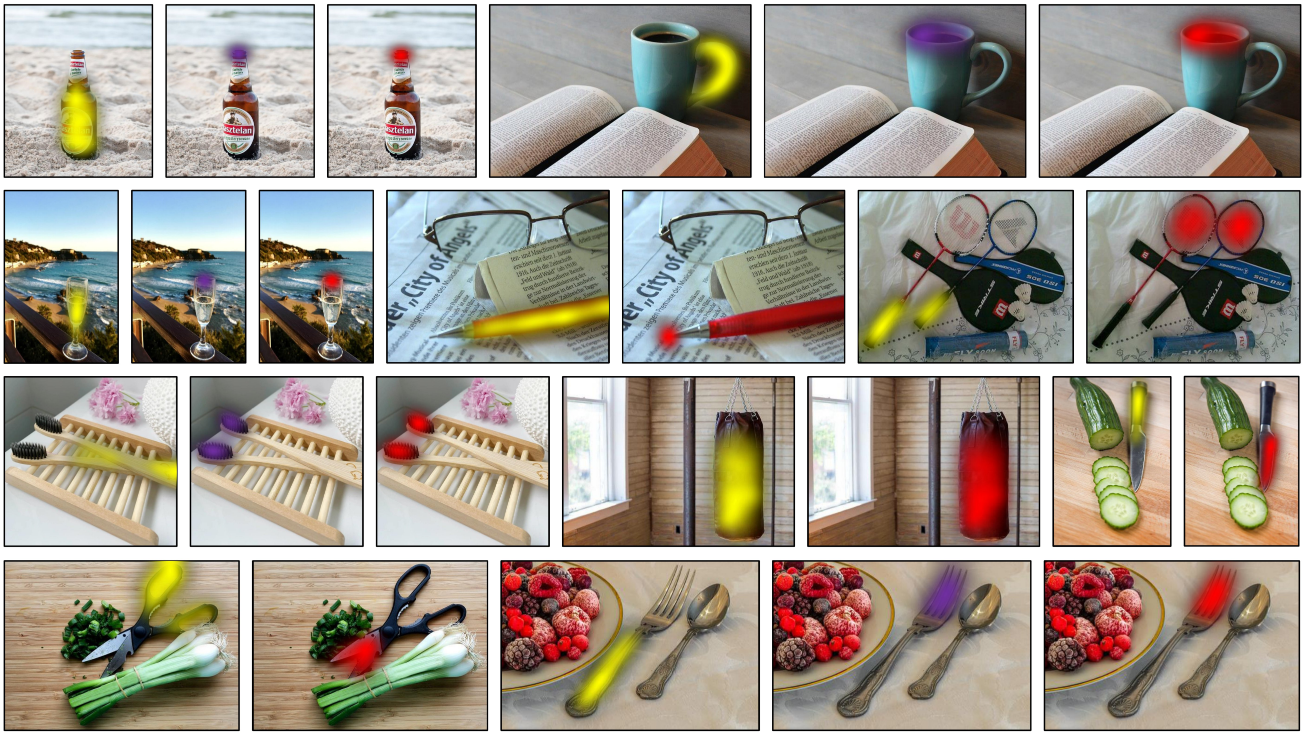


Figure 2. **Non-interactive image examples.** Some examples of non-interactive images and annotations from the CAL dataset.

regions. For the different body part contact regions in the interactive image and the corresponding affordance regions in the non-interactive image, we refer to previous works on affordance/saliency dataset annotation [1, 2, 7, 12] and use the heatmap representation for annotation. Compared to assigning an accurate label to each pixel, heatmap annotation is more descriptive of "action possibilities" (*i.e.*affordance). During the mask generation process, we apply Gaussian blur to all points and normalize them to obtain the corresponding heatmap. Some images and annotations are shown in Fig. 1 and Fig. 2.

A.3. Dataset Division

To comprehensively evaluate the performance of different affordance learning methods and their ability to generalize to unseen interactions/objects, two different divisions, *i.e.*, **Seen** and **Unseen**, are provided. The affordance categories of the training and test sets in **Seen** overlap, and we split the interactive and non-interactive images according to $7:3$, respectively. Then we randomly pair the interactive and non-interactive images in the test set as input and obtain $4,484$ pairs of inputs during testing. For the **Unseen** setting, we choose the categories of "Brush with", "Cut", "Drink with", "Lie on", "Lift" and "Stick" as the test set and the remaining ones as the training set, ensuring that the test set corresponds to human-object interactions covering all body parts. Similarly to **Seen** setting, we randomly select pairs of interactive and non-interactive images belonging to the same affordance category, resulting in a total of $3,297$ testing input pairs.

B. Benchmark

B.1. Metrics

Most affordance learning methods [4, 15, 18, 20] segment precise affordance regions. Thus the metrics they employ are not suitable for evaluating the accuracy of the heatmap of objects' interactable regions. Referring to previous works related to predicting heatmap [2, 7, 13, 16, 19], we choose **KLD [2]**, **SIM [25]** and **NSS [21]** to evaluate the quality of the predicted heatmap in terms of distribution differences, similarity, and correlation.

- **K**ullback-**L**eibler **D**ivergence (**KLD**) **[2]** measures the distribution difference between the prediction (P) and the ground truth (Q). It is computed as follows:

$$KLD\left(P,Q^{D}\right)=\sum_{i}Q_{i}^{D}log\left(\epsilon+\frac{Q_{i}^{D}}{\epsilon+P_{i}}\right),\tag{1}$$

where ϵ is a regularization constant.

- **Sim**ilarity (**SIM**) **[25]** measures the similarity between the prediction map (P) and the ground truth map (Q^D). It is computed as follows:

$$SIM\left(P,Q^{D}\right)=\sum_{i}min\left(P_{i},Q_{i}^{D}\right),\tag{2}$$

where $\sum_i P_i = \sum_i Q_i^D = 1$.

Table 1. **The dimensions, domains of definition, and meanings of the symbols used in the proposed approach.**

	Dimensions	Domains	Meanings
$\boldsymbol{I}_{in}$	$3\times 224\times 224$	[-1,1]	Interactive image
$\boldsymbol{I}_{non}$	$3\times 224\times 224$	[-1,1]	Non-interactive image
$\boldsymbol{P}$	17×2	[0,1]	Human pose
$\boldsymbol{X}_{in_i}$	$c_i\times h_i\times w_i$	$[-\infty,+\infty]$	Interactive feature
$\boldsymbol{X}_{non_i}$	$c_i\times h_i\times w_i$	$[-\infty,+\infty]$	Non-interactive feature
$\boldsymbol{X}_c^i$	$c\times l$	$[-\infty,+\infty]$	Cross-branch tokens
$\boldsymbol{X}_{in}^m/\boldsymbol{X}_{non}^m$	$c\times hw$	$[-\infty,+\infty]$	The IFE module outputs
$\hat{\boldsymbol{F}}_{in}/\hat{\boldsymbol{F}}_{non}$	$c^{'}\times h_1\times w_1$	$[-\infty,+\infty]$	High resolution features
$\boldsymbol{D}_{in}$	$N_{cls}\times h_1\times w_1$	$[0,1]$	Contact region prediction
$\boldsymbol{X}_P$	$17\times c^{'}$	$[-\infty,+\infty]$	Pose feature
$\boldsymbol{H}_j$	$c^{'}\times h_1\times w_1$	$[-\infty,+\infty]$	Contact region feature
$\bar{\boldsymbol{H}}_j$	$c^{'}\times h_1\times w_1$	$[-\infty,+\infty]$	Interactive affinity
$\boldsymbol{D}_{non}$	$N_{cls}\times h_1\times w_1$	$[0,1]$	Non-interaction prediction

- **N**ormalized **S**canpath **S**aliency (**NSS**) **[21]** measures the correspondence between the prediction map (P) and the ground truth (Q^D). It is computed as follows:

$$NSS\left(P,Q^{D}\right)=\frac{1}{N}\sum_{i}\hat{P}\times Q_{i}^{D},\tag{3}$$

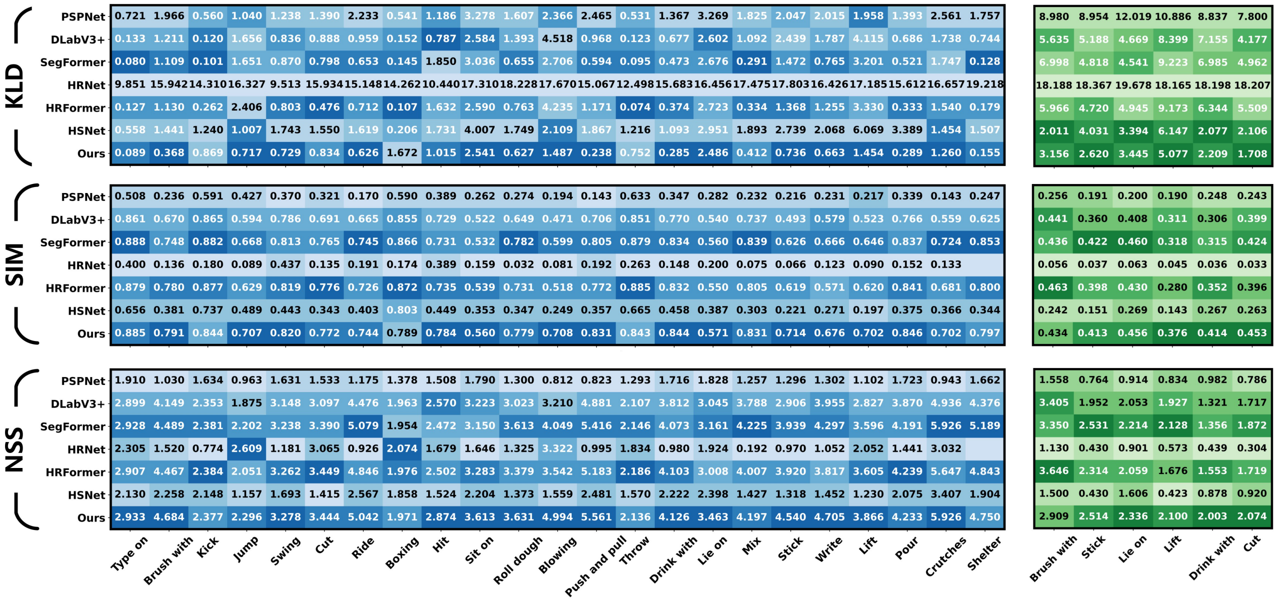
where $N=\sum_i Q_i^D$, $\hat{P}=\frac{P-\mu(P)}{\sigma(P)}$. $\mu\left(P\right)$ and $\sigma\left(P\right)$ are the mean and standard deviation of P, respectively.

B.2. Comparison Methods

To demonstrate the superiority of our model, we select three segmentation models (PSPNet [31], DLabV3+ [3], SegFormer [27]), three human pose estimation models (HRNet [24],ViTPose [28], HRFormer [29]), and one few-shot segmentation model (HSNet [17]) for comparison. Firstly, the task can be considered a segmentation problem for predicting the different interaction regions; therefore, we choose semantic segmentation models. Secondly, human pose keypoints can suppress the effects of diverse interactions and occlusion to obtain a better interactive affinity. The precise prediction of human pose keypoints can more accurately perceive the characteristics of the body part interaction region. Therefore, this paper chooses the human pose estimation models as a comparison. Thirdly, The closest input setting to this task is the few-shot segmentation, and we choose the advanced model for a fairer comparison.

B.3. Implementation Details

We input both interactive and non-interactive images into the network for training for the segmentation and human



		Type on	Brush with	Kick	Jump	Swing	Cut	Ride	Boxing	Hit	Sit on	Roll dough	Blowing	Push and pull	Throw	Drink with	Lie on	Mix	Stick	Write	Lift	Pour	Crutches	Shelter
KLD	PSPNet	0.721	1.966	0.560	1.040	1.238	1.390	2.233	0.541	1.186	3.278	1.607	2.366	2.465	0.531	1.367	3.269	1.825	2.047	2.015	1.958	1.393	2.561	1.757
	DLabV3+	0.133	1.211	0.120	1.656	0.836	0.888	0.959	0.152	0.787	2.584	1.393	4.518	0.968	0.123	0.677	2.602	1.092	2.439	1.787	4.115	0.686	1.738	0.744
	SegFormer	0.080	1.109	0.101	1.651	0.870	0.798	0.653	0.145	1.850	3.036	0.655	2.706	0.594	0.095	0.473	2.676	0.291	1.472	0.765	3.201	0.521	1.747	0.128
	HRNet	9.851	15.942	14.310	16.327	9.513	15.934	15.148	14.262	10.440	17.310	18.228	17.670	15.067	12.498	15.683	16.456	17.475	17.803	16.426	17.185	15.612	16.657	19.218
	HRFormer	0.127	1.130	0.262	2.406	0.803	0.476	0.712	0.107	1.632	2.590	0.763	4.235	1.171	0.074	0.374	2.723	0.334	1.368	1.255	3.330	0.333	1.540	0.179
	HSNet	0.558	1.441	1.240	1.007	1.743	1.550	1.619	0.206	1.731	4.007	1.749	2.109	1.867	1.216	1.093	2.951	1.893	2.739	2.068	6.069	3.389	1.454	1.507
	Ours	0.089	0.368	0.869	0.717	0.729	0.834	0.626	1.672	1.015	2.541	0.627	1.487	0.238	0.752	0.285	2.486	0.412	0.736	0.663	1.454	0.289	1.260	0.155
SIM	PSPNet	0.508	0.236	0.591	0.427	0.370	0.321	0.170	0.590	0.389	0.262	0.274	0.194	0.143	0.633	0.347	0.282	0.232	0.216	0.231	0.217	0.339	0.143	0.247
	DLabV3+	0.861	0.670	0.865	0.594	0.786	0.691	0.665	0.855	0.729	0.522	0.649	0.471	0.706	0.851	0.770	0.540	0.737	0.493	0.579	0.523	0.766	0.559	0.625
	SegFormer	0.888	0.748	0.882	0.668	0.813	0.765	0.745	0.866	0.731	0.532	0.782	0.599	0.805	0.879	0.834	0.560	0.839	0.626	0.666	0.646	0.837	0.724	0.853
	HRNet	0.400	0.136	0.180	0.089	0.437	0.135	0.191	0.174	0.389	0.159	0.032	0.081	0.192	0.263	0.148	0.200	0.075	0.066	0.123	0.090	0.152	0.133	
	HRFormer	0.879	0.780	0.877	0.629	0.819	0.776	0.726	0.872	0.735	0.539	0.731	0.518	0.772	0.885	0.832	0.550	0.805	0.619	0.571	0.620	0.841	0.681	0.800
	HSNet	0.656	0.381	0.737	0.489	0.443	0.343	0.403	0.803	0.449	0.353	0.347	0.249	0.357	0.665	0.458	0.387	0.303	0.221	0.271	0.197	0.375	0.366	0.344
	Ours	0.885	0.791	0.844	0.707	0.820	0.772	0.744	0.789	0.784	0.560	0.779	0.708	0.831	0.843	0.844	0.571	0.831	0.714	0.676	0.702	0.846	0.702	0.797
NSS	PSPNet	1.910	1.030	1.634	0.963	1.631	1.533	1.175	1.378	1.508	1.790	1.300	0.812	0.823	1.293	1.716	1.828	1.257	1.296	1.302	1.102	1.723	0.943	1.662
	DLabV3+	2.899	4.149	2.353	1.875	3.148	3.097	4.476	1.963	2.570	3.223	3.023	3.210	4.881	2.107	3.812	3.045	3.788	2.906	3.955	2.827	3.870	4.936	4.376
	SegFormer	2.928	4.489	2.381	2.202	3.238	3.390	5.079	1.954	2.472	3.150	3.613	4.049	5.416	2.146	4.073	3.161	4.225	3.939	4.297	3.596	4.191	5.926	5.189
	HRNet	2.305	1.520	0.774	2.609	1.181	3.065	0.926	2.074	1.679	1.646	1.325	3.322	0.995	1.834	0.980	1.924	0.192	0.970	1.052	2.052	1.441	3.032	
	HRFormer	2.907	4.467	2.384	2.051	3.262	3.449	4.846	1.976	2.502	3.283	3.379	3.542	5.183	2.186	4.103	3.008	4.007	3.920	3.817	3.605	4.239	5.647	4.843
	HSNet	2.130	2.258	2.148	1.157	1.693	1.415	2.567	1.858	1.524	2.204	1.373	1.559	2.481	1.570	2.222	2.398	1.427	1.318	1.452	1.230	2.075	3.407	1.904
	Ours	2.933	4.684	2.377	2.296	3.278	3.444	5.042	1.971	2.874	3.613	3.631	4.994	5.561	2.136	4.126	3.463	4.197	4.540	4.705	3.866	4.233	5.926	4.750

		Brush with	Stick	Lie on	Lift	Drink with	Cut
KLD	PSPNet	8.980	8.954	12.019	10.886	8.837	7.800
	DLabV3+	5.635	5.188	4.669	8.399	7.155	4.177
	SegFormer	6.998	4.818	4.541	9.223	6.985	4.962
	HRNet	18.188	18.367	19.678	18.165	18.198	18.207
	HRFormer	5.966	4.720	4.945	9.173	6.344	5.509
	HSNet	2.011	4.031	3.394	6.147	2.077	2.106
	Ours	3.156	2.620	3.445	5.077	2.209	1.708
SIM	PSPNet	0.256	0.191	0.200	0.190	0.248	0.243
	DLabV3+	0.441	0.360	0.408	0.311	0.306	0.399
	SegFormer	0.436	0.422	0.460	0.318	0.315	0.424
	HRNet	0.056	0.037	0.063	0.045	0.036	0.033
	HRFormer	0.463	0.398	0.430	0.280	0.352	0.396
	HSNet	0.242	0.151	0.269	0.143	0.267	0.263
	Ours	0.434	0.413	0.456	0.376	0.414	0.453
NSS	PSPNet	1.558	0.764	0.914	0.834	0.982	0.786
	DLabV3+	3.405	1.952	2.053	1.927	1.321	1.717
	SegFormer	3.350	2.531	2.214	2.128	1.356	1.872
	HRNet	1.130	0.430	0.901	0.573	0.439	0.304
	HRFormer	3.646	2.314	2.059	1.676	1.553	1.719
	HSNet	1.500	0.430	1.606	0.423	0.878	0.920
	Ours	2.909	2.514	2.336	2.100	2.003	2.074

Figure 3. **Different Classes.** We measure the KLD, SIM, and NSS metrics for each affordance category, with darker colors representing higher performance. The left and right represent the results at the **Seen** and **Unseen** settings, respectively.

Table 2. **The influence of m.** We investigate the impact of the hyper-parameter m (*i.e.*the number of layers in the IFE block) in the IFE module on model performance.

	m=?	KLD ↓	SIM ↑	NSS ↑
Seen	1	1.275	0.731	3.517
	2	**0.965**	**0.756**	**3.723**
	3	1.024	0.743	3.702
	4	1.126	0.734	3.654
	5	1.217	0.729	3.594
Unseen	1	4.403	0.411	2.127
	2	**2.823**	**0.430**	**2.303**
	3	3.101	0.416	2.176
	4	3.471	0.399	2.066
	5	5.181	0.384	1.935

pose estimation models. In the testing process, we only input non-interactive images to predict the corresponding affordance regions. All models are trained with the default parameters, and the input images are 224 × 224, using the same data augmentation. The final output is fed through a Sigmoid activation function, and the loss is calculated using a binary cross-entropy loss. For the few-shot segmentation model, we take the interactive image along with a body part contact region as a group of inputs for the support branch (*i.e.*each pair of interactive and non-interactive image pairs corresponds to N_{cls} groups of inputs) and predict the corresponding interactive regions in the non-interactive images.

Table 3. **The influence of l.** We investigate the impact of the hyper-parameter l (*i.e.*the number of cross-branch tokens) in the IFE module on model performance.

	l=?	KLD ↓	SIM ↑	NSS ↑
Seen	2	1.938	0.676	3.434
	4	1.252	0.731	3.585
	8	1.150	0.742	3.653
	16	**0.965**	**0.756**	**3.723**
	32	1.164	0.735	3.671
Unseen	2	5.060	0.396	2.079
	4	4.216	0.409	2.134
	8	3.045	0.409	2.181
	16	**2.823**	**0.430**	**2.303**
	32	3.972	0.418	2.208

Furthermore, Table 1 shows the dimensions, definition domains, and meanings of the corresponding symbols in the method proposed in this paper.

C. Experiments

C.1. Different Classes

We explore the performance of different methods on different affordance classes. Fig. 3 shows their results on each category in the KLD, SIM, and NSS metrics, in which darker colors indicate that the model performs better on individual categories. In the **Seen** setting, our method outperforms all other methods in most metrics, while in the

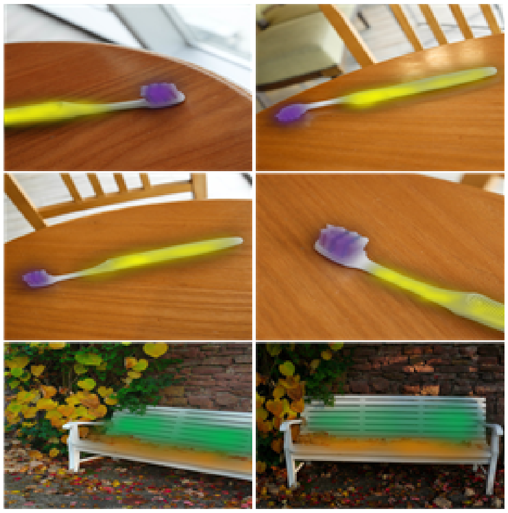


Figure 4. **Different Views.**

`Unseen` setting, the results are either best or second best in most metrics.

C.2. Different Hyper-parameters

Table 2 and Table 3 show the impact of the hyper-parameters m and l in the IFE module on the model's performance, where $l = 16$ and $m = 2$ are the default settings in the paper. The influence at the `Unseen` setting is slightly more apparent, where the number of cross-branch tokens greatly affects the model performance.

C.3. Different Views

Our dataset contains interactive and non-interactive images from different views, enabling the trained model to adapt to viewpoint changes. Besides, the IFE module can establish contextual links between different body contact regions, which helps the model counteract the effects of viewpoint changes. The non-interactive images from different viewpoints are shown in the figure, indicating the robustness under viewpoint changes. Further, we will collect a test set containing different views of the same object to evaluate the model's robustness to viewpoint changes.

C.4. More Visualization Results

Fig. 5 and Fig. 6 show the prediction results of all methods in the `Seen` and `Unseen` settings, respectively.

E. Potential Applications

- **Human-Object Interaction.** Our method is able to predict the interactive affinity of each body part and object local regions from the human-object interaction images, which can facilitate the network to establish the connection between different body parts [6, 9] and infer the corresponding interactions, as well as reducing the interference of redundant negative pairs to improve the accuracy of prediction. [14, 26].
- **AR/VR** Furthermore, our method can be used in AR/VR [5, 23], *i.e.*, for humans in different physical scenes, our method is able to extract the body parts corresponding contact regions during their interactions with the environment and jointly map them to a common scene rendered by the VR device, facilitating the device to generate scenes with human interactions in the virtual world [10, 11, 22].

References

[1] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 3

[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. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3, 6, 7

[4] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 3

[5] Ersin Dincelli and Alper Yayla. Immersive virtual reality in the age of the metaverse: A hybrid-narrative review based on the technology affordance perspective. *The Journal of Strategic Information Systems*, 31(2):101717, 2022. 5

[6] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–67, 2018. 5

[7] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018. 3

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 1

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 5

[10] Vladimir Guzov, Torsten Sattler, and Gerard Pons-Moll. Visually plausible human-object interaction capture from wearable sensors. *arXiv preprint arXiv:2205.02830*, 2022. 5

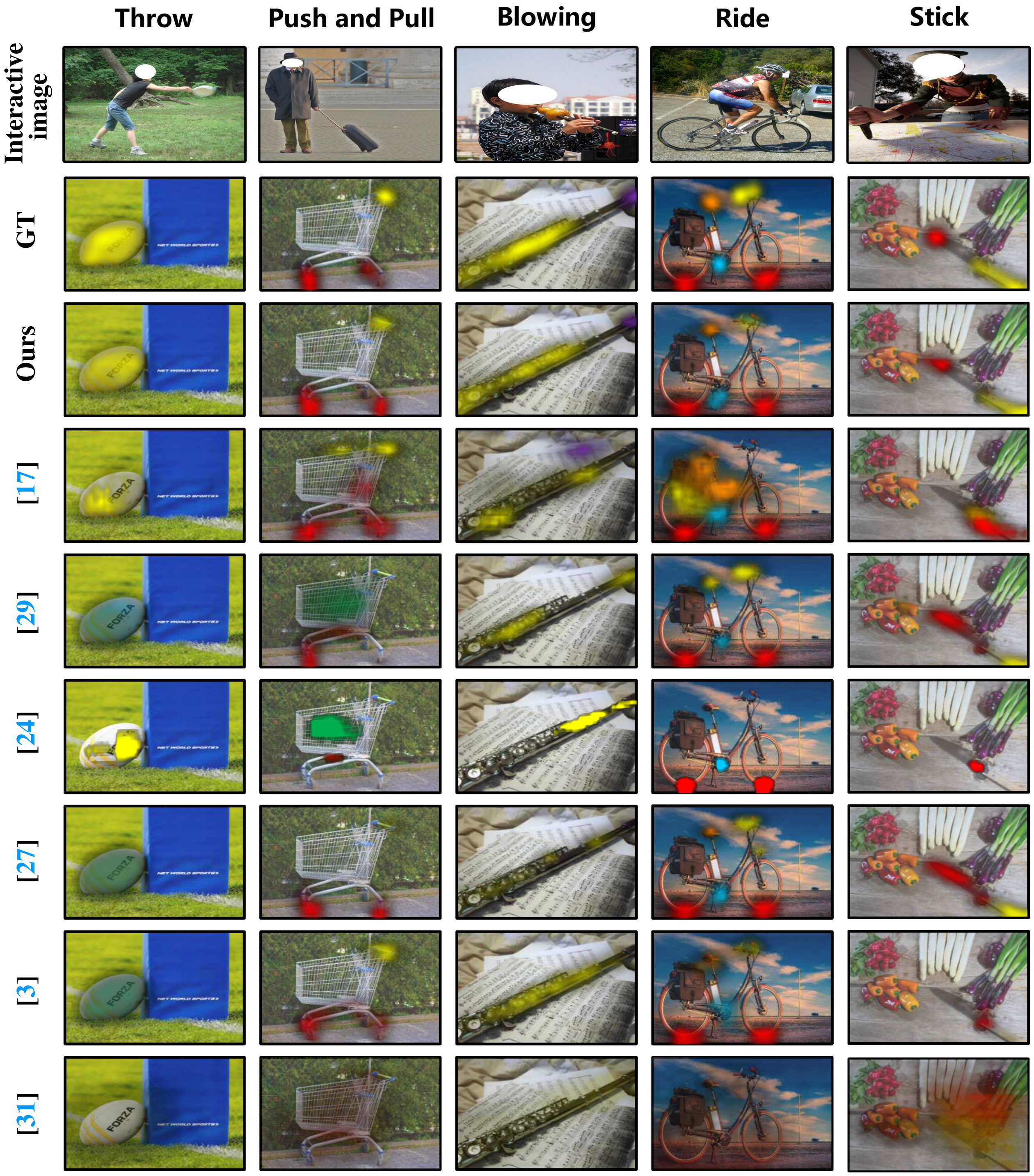


Figure 5. **Visualization results.** Prediction results for each model under the `Seen` setting.

[11] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 5

[12] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 3

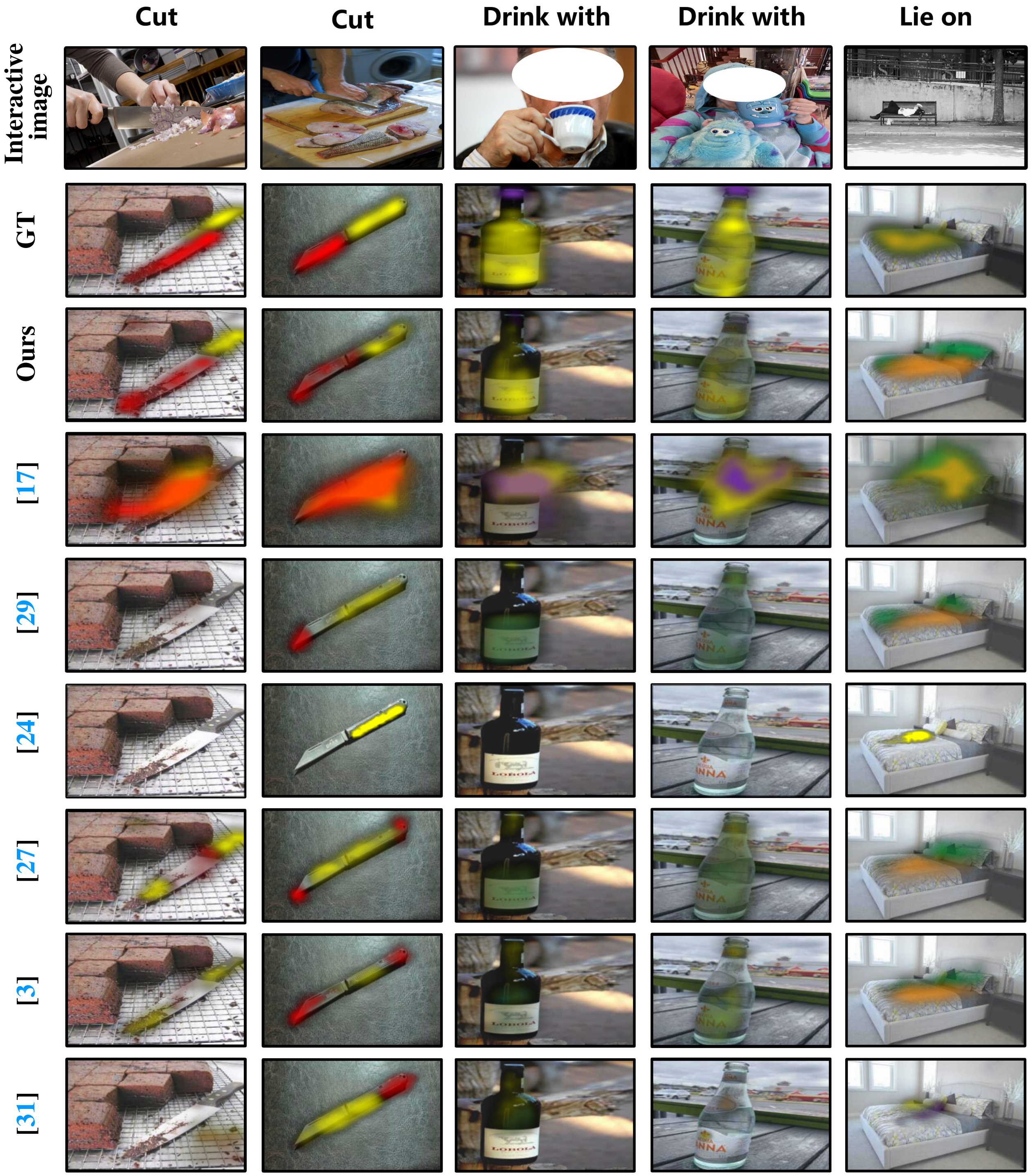


Figure 6. **Visualization results.** Prediction results for each model under the `Unseen` setting.

[13] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. *arXiv preprint arXiv:2204.01696*, 2022. 3

[14] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages

20113–20122, 2022. 5

[15] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. *arXiv preprint arXiv:2106.14747*, 2021. 3

[16] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2252–2261, 2022. 1, 3

[17] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 7

[18] 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

[19] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 3

[20] 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

[21] 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

[22] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 5

[23] Dong-Hee Shin. The role of affordance in the experience of virtual reality learning: Technological and affective affordances in virtual reality. *Telematics and Informatics*, 34(8):1826–1836, 2017. 5

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3, 6, 7

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

[26] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *European Conference on Computer Vision*, pages 121–136. Springer, 2022. 5

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 6, 7

[28] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 3

[29] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. 2021. 3, 6, 7

[30] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500, 2022. 1

[31] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3, 6, 7