Leverage Interactive Affinity for Affordance Learning

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

Abstract

Perceiving potential "action possibilities" (i.e., affordance) regions of images and learning interactive functionalities of objects from human demonstration is a challenging task due to the diversity of human-object interactions. Prevailing affordance learning algorithms often adopt the label assignment paradigm and presume that there is a unique relationship between functional region and affordance label, yielding poor performance when adapting to unseen environments with large appearance variations. In this paper, we propose to leverage interactive affinity for affordance learning, i.e.extracting interactive affinity from human-object interaction and transferring it to non-interactive objects. Interactive affinity, which represents the contacts between different parts of the human body and local regions of the target object, can provide inherent cues of interconnectivity between humans and objects, thereby reducing the ambiguity of the perceived action possibilities. Specifically, we propose a pose-aided interactive affinity learning framework that exploits human pose to guide the network to learn the interactive affinity from human-object interactions. Particularly, a keypoint heuristic perception (KHP) scheme is devised to exploit the keypoint association of human pose to alleviate the uncertainties due to interaction diversities and contact occlusions. Besides, a contact-driven affordance learning (CAL) dataset is constructed by collecting and labeling over $5,000$ *images. Experimental results demonstrate that our method outperforms the representative models regarding objective metrics and visual quality. Code and dataset: github.com/lhc1224/PIAL-Net.*

1. Introduction

The objective of affordance learning is to locate the "action possibilities" regions of an object [15, 18]. For an intelligent agent, it is vital to perceive not only the object semantics but also how to interact with various objects' local regions. Perceiving and reasoning about the object's interactable regions is a critical capability for embodied intelligent systems to interact with the environment actively, distinct from passive perception systems [3, 38, 39, 44]. Moreover, affordance learning has a wide range of applications in fields such as action recognition [13, 24, 43], scene understanding [9, 69], human-robot interaction [51, 63], autonomous driving [7] and VR/AR [50, 53].

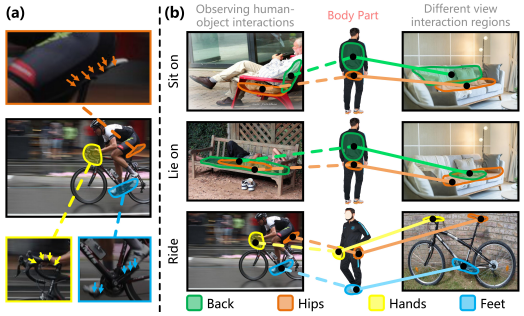

Figure 1. **(a)** Interaction affinity refers to the contact between different parts of the human body and the local regions of a target object. **(b)** The interactive affinity provides rich cues to guide the model to acquire invariant features of the object's local regions interacting with the body part, thus counteracting the multiple possibilities caused by diverse interactions.

Affordance is a dynamic property closely related to humans and the environment [18]. Previous works [11, 37, 40, 46] focus on establishing mapping relationships between appearances and labels for affordance learning. However, they neglect the multiple possibilities of affordance brought about by changes in the environment and actors, leading to an incorrect perception. Recent studies [39, 48] utilize reinforcement learning to allow intelligent agents to perceive the environment through numerous interactions in simulated/actual scenarios. Such approaches are mainly limited by their high cost and struggle to generalize to

*Corresponding author. ‡ Equal contributions.

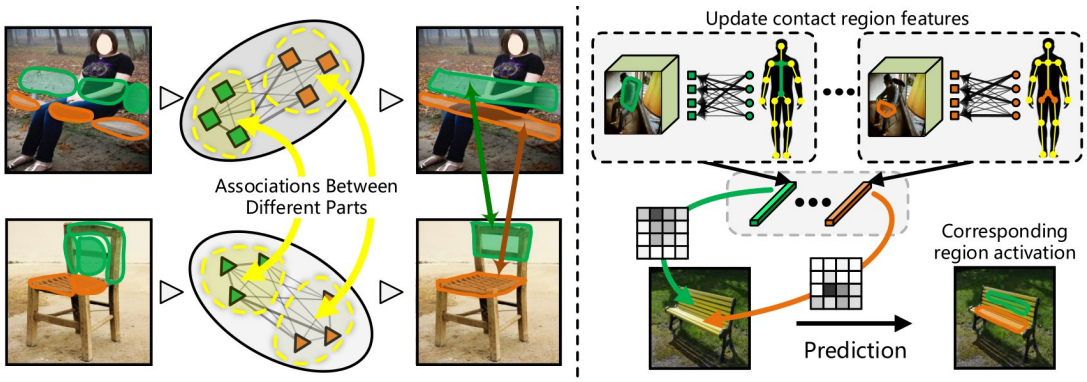

Figure 2. **Motivation. (a)** This paper explores the associations of interactable regions between diverse images by considering the context of contact regions with different body parts. **(b)** This paper considers leveraging the connection of human pose keypoints to alleviate the uncertainties due to interaction diversities and contact occlusions.

unseen scenarios [58]. To this end, researchers consider learning from human demonstration in an action-free manner [14, 29, 30, 38]. Nonetheless, they only roughly segment the whole object/interaction regions in a general way, which is still challenging to understand how the object is used. The multiple possibilities due to different local regions interacting with humans in various ways are not fully resolved. In this paper, we propose to leverage interactive affinity for affordance learning, *i.e.*extracting interactive affinity from human-object interaction and transferring it to non-interactive objects. The interactive affinity (Fig. 1 (a)) denotes the contacts between different human body parts and objects' local regions, which can provide inherent cues of interconnectivity between humans and objects, thereby reducing the ambiguity of the perceived action possibilities (Fig. 1 (b)).

However, it faces the challenges of interaction diversities and contact occlusions, leading to difficulties in acquiring a good interactive affinity representation. The human pose is independent of background, and the same interaction corresponds to approximately similar poses. Thus, it makes sense to use the association between pose keypoints to overcome the difficulty of obtaining interactive affinity representations. Moreover, it is challenging to transfer the interactive affinity to non-interactive object images due to variations in views, scales, and appearances. The context between the different body part contact regions (Fig. 2 (a)) provides the model with the possibility to explore the associations between the interactable regions of the various images to counteract transfer difficulties.

In this paper, we present a pose-aided interactive affinity learning framework. First, an **I**nteractive **F**eature **E**nhancement (**IFE**) module is introduced to explore the connections between different interactable regions of the images. Then, a **K**eypoint **H**euristic **P**erception (**KHP**) scheme is devised to mine the interactive affinity representation from interaction and transfer it to non-interactive objects. Specifically, the IFE module leverages the transformer to fully extract global contextual cues by exploiting the common relationships between their local interactable regions (Fig. 2 (a)). Then, they are used to establish associations between the object interactable regions in different images. Subsequently, the KHP scheme exploits the correlation between the human body keypoints and the contact region to guide the network to mine the object's local invariant features interacting with the body parts (Fig. 2 (b)).

Although the numerous related datasets [9, 10, 19, 30, 36, 57, 67] that emerged during the development of affordance learning, there is still a lack of relevant datasets suited for leveraging interactive affinity. To carry out a thorough study, this paper constructs an **C**ontact-driven **A**ffordance **L**earning (**CAL**) dataset, consisting of $5,258$ images from 23 affordance and 47 object categories. We conduct contrastive studies on the CAL dataset against six representative models in several related fields. Experimental results validate the effectiveness of our method in solving the multiple possibilities of affordance.

Contributions: 1) We propose leveraging interactive affinity for affordance learning and establishing a CAL benchmark to facilitate the study of obtaining interactive affinity to counteract the multiple possibilities of affordance. **2)** We propose a pose-aided interactive affinity learning framework that exploits pose data to guide the network to mine the interactive affinity of body parts and object regions from human-object interaction. **3)** Experiments on the CAL dataset demonstrate that our model outperforms state-of-the-art methods and can serve as a strong baseline for future affordance learning research.

2. Related Work

2.1. Affordance Learning

Affordance learning has attracted extensive attention in recent years due to its immense value in many fields

such as robotics, autonomous driving, embodied AI, *etc* [18,20,32,63]. Early works [10,11,34,35,37,40,46,47,55] mainly adopt the label assignment paradigm and presume that there is a unique relationship between functional region and affordance label. Nevertheless, it is hard to deal with the issue of the multiple possibilities due to environmental and operator changes. Recent studies consider using reinforcement learning by setting different reward functions to empower the agent to actively interact with the environment to acquire the ability to perceive affordance [26, 39, 48]. These methods involve extensive interactions and are costly in real scenarios, while there is still a large domain gap between the simulation and the real scenario. They also have the limitation of being poorly adaptive to unseen scenarios. Some other works consider learning the object's affordance from the human demonstration in an action-free manner [14, 29, 30, 38, 67], extracting interactions from the images/videos and transferring the human action intentions implied within them to the new unseen object, thus achieving perception and generalization. However, they only detect/segment the object as a whole or the interactable regions in a general way. They do not perceive how the object's local regions are used and have not fully resolved the multiple possibilities issue. In contrast to the above works, this paper considers using the inherent cues of interconnectivity between humans and objects to reduce the ambiguity of the perceived action possibilities.

2.2. Body Part Contact Learning

The contact between the body part and the object is an important clue for 3D reconstruction and human motion forecasting [2, 17, 31, 45, 49, 60, 65]. Independent estimation of human pose and object during 3D scene reconstruction can lead to incorrect body-scene interpolation and body floating. Human-scene contacts can provide reliable boundary conditions for improved 3D pose estimation and localization. Shimada et al. [49] use body-scene contacts to guide 3D human capture. Bhatnagar et al. [2] propose to jointly track humans, objects, and contacts along with collecting a large-scale BEHAVE dataset containing human models, objects, and contact annotations. Mao et al. [31] introduce distance-based contact maps as an explicit constraint for human motion forecasting. Yang et al. [65] propose to model hand-object interaction by explicitly representing the contact using the Contact Potential Field (CPF). In contrast to [65], this paper utilizes the interactive affinity to mine the interconnectivity between humans and objects, helping reduce the ambiguity in affordance learning.

3. Method

Given a human-object interaction image I_{in} with a corresponding human pose P, and a non-interactive image I_{non}, we aim to extract the affordance affinity representation between the human body part and the object local region from I_{in} and transfer it to I_{non} to predict the corresponding interactable region. The pose-aided interactive affinity learning framework is shown in Fig. 3. It first extracts features through a transformer [59] backbone to obtain $\mathbb{X}_{in} = \{\boldsymbol{X}_{in_i}, i \in [1,4]\}$ and $\mathbb{X}_{non} = \{\boldsymbol{X}_{non_i}, i \in [1,4]\}$, respectively ($i$ indexes the block of the backbone). Then, an interactive feature enhancement (IFE) module (Sec. 3.1) is introduced to establish a correlation between the interactable features from diverse images. Finally, a keypoint heuristic perception (KHP) scheme (Sec. 3.2) uses the association of human pose to mine the interactive affinity representation and then transfers it to the non-interactive image to predict the corresponding interactive region.

3.1. Interactive Feature Enhancement Module

The IFE module is shown in Fig. 3. Due to various views, scales, and appearances existing in the I_{in} and I_{non}, it is difficult to transfer the interactive affinity to the non-interactive branch. Therefore, we consider narrowing the discrepancy between the two branches by establishing the connection between the interactable regions before extracting the interactive affinity representation from human-object interaction. Inspired by the advantages of the transformer in modeling long-range global contextual information and its scalability of dealing with varied-length sequence [12, 22, 62], the IFE module adopts the transformer to associate cross-branch interactable regions. The cross transformer ($\boldsymbol{Z}_1^{i+1} = \mathrm{CT}(\boldsymbol{Z}_1^i, \boldsymbol{Z}_2^i)$) is computed as:

$$\boldsymbol{Y}_1^i = \mathrm{MCA}(\mathrm{LN}(\boldsymbol{Z}_1^i), \mathrm{LN}(\boldsymbol{Z}_2^i)) + \boldsymbol{Z}_1^i, \tag{1}$$

$$\boldsymbol{Z}_1^{i+1} = \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Y}_1^i)) + \boldsymbol{Y}_1^i, \tag{2}$$

where $\mathrm{MCA}()$ denotes the dot-production attention [56], where query, key, and value go through different linear layers: $\mathrm{MCA}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{Attention}(\boldsymbol{W}^Q\boldsymbol{X}, \boldsymbol{W}^K\boldsymbol{Y}, \boldsymbol{W}^V\boldsymbol{Y})$. The IFE module initializes a series of cross-branch tokens, which aim to aggregate the two branches' global contextual cues. Specifically, the cross-branch token $\boldsymbol{X}_c^0 \in \mathbb{R}^{l\times c}$ (l and c denote the number of tokens and channels, respectively) first extracts the contextual information of the two branches separately. It then extracts the contextual information common to $\boldsymbol{X}_{in}$ and $\boldsymbol{X}_{non}$ by mining the co-contextual features of their interactable regions. $\boldsymbol{X}_c^i$ is updated as:

$$\boldsymbol{X}_c^i = \mathrm{CT}(\boldsymbol{X}_c^i, [\boldsymbol{X}_{in}^i, \boldsymbol{X}_c^i]), \boldsymbol{X}_c^i = \mathrm{CT}(\boldsymbol{X}_c^i, [\boldsymbol{X}_{non}^i, \boldsymbol{X}_c^i]), \tag{3}$$

$$\boldsymbol{X}_c^{i+1} = \mathrm{CT}(\boldsymbol{X}_c^i, [\boldsymbol{X}_{in}^i, \boldsymbol{X}_{non}^i, \boldsymbol{X}_c^i]), \tag{4}$$

where $[\cdot,\cdot]$ represents the operation of concatenating two tensors. $\boldsymbol{X}_{in}^i$ and $\boldsymbol{X}_{non}^i$ ($i > 0$) are the two branches feature sequences of the i-th IFE block, respectively. $\boldsymbol{X}_{in_4}$ and $\boldsymbol{X}_{non_4}$ are reshaped to $\mathbb{R}^{wh\times c}$ (w and h are the width and height of the feature map, respectively) to obtain $\boldsymbol{X}_{in}^0$ and $\boldsymbol{X}_{non}^0$, respectively. After updating the cross-branch

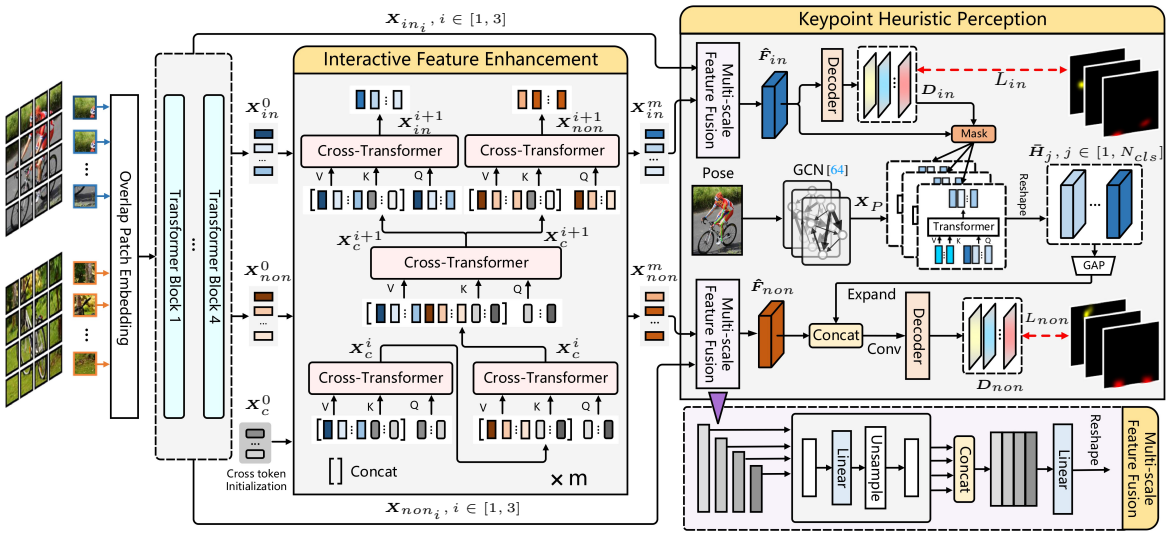

Figure 3. **Overview of the proposed pose-aided interactive affinity learning framework.** Our model mainly consists of an interactive feature enhancement (IFE) module and a keypoint heuristic perception (KHP) scheme.

tokens, the IFE module then uses $\boldsymbol{X}_c^{i+i}$ to update the interactive and non-interactive features, respectively, enabling the two branches to effectively strengthen the interaction-related features from a different branch and reduce the variation. The update procedure is as follows:

$$\boldsymbol{X}_{in}^{i+1} = \mathrm{CT}(\boldsymbol{X}_{in}^{i}, [\boldsymbol{X}_{in}^{i}, \boldsymbol{X}_{c}^{i+1}]), \tag{5}$$

$$\boldsymbol{X}_{non}^{i+1} = \mathrm{CT}(\boldsymbol{X}_{non}^{i}, [\boldsymbol{X}_{non}^{i}, \boldsymbol{X}_{c}^{i+1}]). \tag{6}$$

After passing through m IFE blocks, we obtain the output features $\boldsymbol{X}_{in}^{m}$ and $\boldsymbol{X}_{non}^{m}$, respectively, which integrate cross-branch interactable regions information.

3.2. Keypoint Heuristic Perception Scheme

The KHP scheme is shown in Fig. 3. First, the IFE module's outputs ($\boldsymbol{X}_{in}^{m}$ / $\boldsymbol{X}_{non}^{m}$) are sent to a multi-scale feature fusion layer to fuse with the shallow features extracted from the backbone, to obtain the higher resolution features [59]:

$$\boldsymbol{F}_{in_i} = \mathrm{Upsample}(\mathrm{Linear}(c, c^{'})(\boldsymbol{X}_{in}^{m})), \quad i = 4, \tag{7}$$

$$\boldsymbol{F}_{in_i} = \mathrm{Upsample}(\mathrm{Linear}(c_i, c^{'})(\boldsymbol{X}_{in_i})), \quad i \in [1, 3], \tag{8}$$

$$\hat{\boldsymbol{F}}_{in} = \mathrm{Linear}(4c^{'}, c^{'})(\mathrm{Concat}(\boldsymbol{F}_{in_i})), \quad i \in [1, 4], \tag{9}$$

where $\mathrm{Linear}(c_i, c_o)$ denotes a linear layer with inputs C_i and outputs c_o, respectively. $\hat{\boldsymbol{F}}_{non}$ is calculated similar to $\hat{\boldsymbol{F}}_{in}$. The $\hat{\boldsymbol{F}}_{in}$ is reshaped to 2D feature map and sent to a decoder $\mathrm{Decoder}(\cdot)$ to obtain the contact prediction: $\boldsymbol{D}_{in} = \mathrm{Decoder}(\hat{\boldsymbol{F}}_{in})$. $\boldsymbol{D}_{in}$ contains N_{cls} (N_{cls} is the number of body parts) channels, and the feature map at each channel represents the regions where different body parts interact with the object. The diversity of human body styles, clothing, interaction postures, and contact occlusion make it difficult to obtain an accurate interactive affinity representation of body parts and object regions. The KHP scheme considers using the human pose to assist the network in suppressing irrelevant background regions and establishing a connection between body parts and contact regions. We first use a graph convolutional network (GCN) [1, 23, 64] to extract the features of pose P:

$$\boldsymbol{X}_P = \boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A} \otimes \boldsymbol{M} + \boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{W}_{gcn}, \tag{10}$$

where $\boldsymbol{P}$ indicates the coordinates of the keypoints' locations, $\boldsymbol{A}$ is the adjacency matrix and $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{\Lambda}_{ii} = \sum_i(\boldsymbol{A}_{ij} + \boldsymbol{I}_{ij})$, $\boldsymbol{W}_{gcn}$ is the weight matrix. $\boldsymbol{M}$ is a learnable parameter matrix measuring the importance between edges, which has the same dimension as $\boldsymbol{A}$ and is initialized to all 1 [64]. Subsequently, we use the pose keypoint features corresponding to the body parts to guide the model to extract the corresponding interactive affinity representation from the contact region:

$$\boldsymbol{H}_j = \hat{\boldsymbol{F}}_{in} \otimes \boldsymbol{P}_{in_j}, \quad \bar{\boldsymbol{H}}_j = \mathrm{CT}(\boldsymbol{H}_j, \boldsymbol{X}_{P_j}) \tag{11}$$

where $\boldsymbol{X}_{P_j}$ denotes the feature of the pose keypoint corresponding to the j-th ($j \in [1, N_{cls}]$) body part. Following this, we transfer the interactive affinity representation to the non-interactive object. We feed the interactive affinity features $\bar{\boldsymbol{H}}_j$ into the global average pooling layer (GAP) to obtain the feature representation $\boldsymbol{f}_j$:$\boldsymbol{f}_j = \mathrm{GAP}(\bar{\boldsymbol{H}}_j)$. Subsequently, $\boldsymbol{f}_j$ is expanded to the size of $\hat{\boldsymbol{F}}_{non}$, concatenated with $\hat{\boldsymbol{F}}_{non}$ and fed into a convolution layer to obtain the transferred features $\boldsymbol{K}_j$:

$$\boldsymbol{K}_j = \mathrm{Conv}(\mathrm{Concat}(\hat{\boldsymbol{F}}_{non}, \mathrm{Expand}(\boldsymbol{f}_j))). \tag{12}$$

Finally, $\boldsymbol{K}_j$ is concatenated together and fed into a decoder to obtain the prediction of the non-interactive branch: $\boldsymbol{D}_{non} = \mathrm{Decoder}(\mathrm{Concat}(\boldsymbol{K}_j)), j \in [1, N_{cls}]$. During

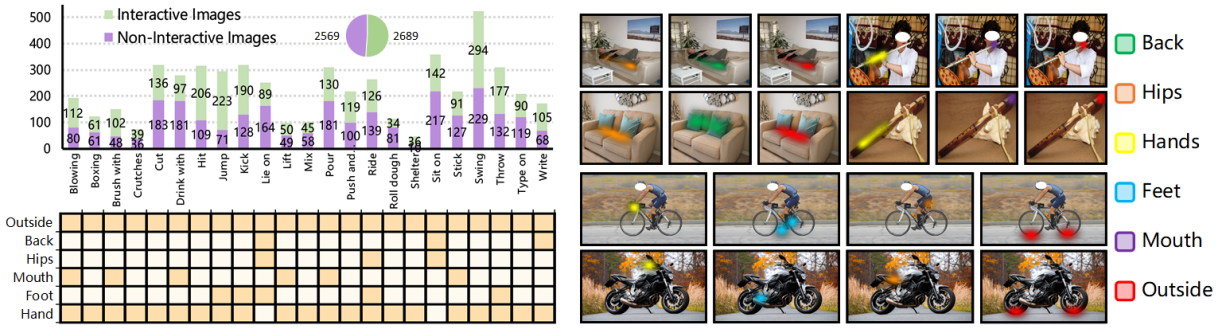

Figure 4. **Some examples and properties of Contact-driven Affordance Learning (CAL) dataset.** **(a)** Statistics on the quantity of interactive and non-interactive images in each affordance category. **(b)** Confusion matrix for each affordance category interacting with body parts. **(c)** Some examples of interactive and non-interactive images and annotations in the dataset.

training, the total loss is defined as: $L_{total} = \lambda_1 L_{in} + \lambda_2 L_{non}$, where L_{in} and L_{non} are the binary cross-entropy losses for the two branches, respectively. λ_1 and λ_2 are the loss weight parameters and are both set to 1.

4. Experiment

4.1. Dataset

Dataset collection. To fill the gap of lacking a suitable dataset, we select $5,265$ images from the PAD/PADv2 [29, 67] and AGD20K [30] to compose the **C**ontact-driven **A**ffordance **L**earning (**CAL**) dataset. Some examples are shown in Fig. 4 (c). We refer to [6, 27] and choose the commonly used 23 affordance and 47 object categories. We retrieve images according to affordance categories, then manually filter and classify them. Finally, the dataset contains $2,689$ interactive images and $2,569$ non-interactive images.

Dataset annotation. We consider the interaction of six different body parts: "Hands", "Feet", "Mouth", "Hips", "Back" and "Outside' (the "Outside" represents the region where the object contacts the outside world during the interaction, *e.g.*, the wheel-ground contact region during riding). These categories cover almost all regions where humans interact with objects daily and thoroughly describe how various object regions are used. Since affordance learning recognizes "action possibilities" regions of the object, the heatmap is appropriate for describing the possibility of interactions. We refer to the previous annotation works [4, 5, 14, 21] and choose to annotate local regions of the image with different densities of points. Each point is first Gaussian blurred and normalized during generating the soft mask to obtain the final interaction correspondence map. Some annotated examples are shown in Fig. 4 (c).

Statistical analysis. We conduct some statistical analysis to get a deeper insight into the dataset. Fig. 4 (a) shows the number of interactive and non-interactive images for each affordance category. The number of interactive and non-interactive images is comparable, but data imbalance still exists between different categories. Fig. 4 (b) shows the confusion matrix for the affordance class with body part contact, where multiple contact regions exist for each category, implying the challenge of affordance learning.

4.2. Benchmark Settings

To provide a comprehensive evaluation, three commonly used metrics **K**ullback-**L**eibler **D**ivergence (**KLD**) [5], **SIM**ilarity (**SIM**) [54], and **N**ormalized **S**canpath **S**aliency (**NSS**) [42] are chosen, see supplementary material for more details. Our model is implemented in PyTorch and trained with the AdamW [28] optimizer. The input images are cropped to 224×224. We train the model for $60,000$ iterations on a single NVIDIA 3090ti GPU with an initial learning rate of $6e$-5 and a batch size of 16. The hyper-parameters l and m in the IFE module are set to 16 and 2, respectively. Furthermore, we provide two different settings to evaluate model learning and generalization capabilities: **1)** **`Seen`**, *i.e.*, the training set and the test set contain the same affordance/object categories, and **2)** **`Unseen`**, *i.e.*, the affordance/object categories in the training set and the test set do not overlap. The test set in **`Seen`** setting contains $4,484$ interactive and non-interactive pairs, while the **`Unseen`** setting contains $3,297$ pairs. The semantic segmentation method is chosen to compare the advantages of the methods in overcoming the multiple possibilities. We also choose the human pose estimation methods for comparing the interactive affinity for affordance learning approach. Besides, the few-shot segmentation model is also chosen for a fair comparison. In summary, we select 3 segmentation (DeepLabV3+ [8], PSPNet [68], Segformer [59]), 3 pose estimation (HRNet [52], ViTPose [61], HRFormer [66]), and one few-shot segmentation (HSNet [33]) methods from the relevant fields for comparison.

4.3. Quantitative and Qualitative Comparisons

The experimental results of different methods on the CAL dataset are shown in Table 1. Our method outper-

Table 1. **The results of different methods on the CAL dataset.** The best results are in **bold**. **Seen** means that the training and test sets contain the same affordance/object categories, and **Unseen** means that the affordance/objects in the training and test sets do not overlap. "♢" , "♣", and "♠" represent segmentation, human pose estimation, and few-shot segmentation models, respectively. The ⋄ defines the relative improvement of our method over other methods.

Method		Seen			Unseen			params
		KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	(M)
PSPNet [68]		1.738 ⋄44.5%	0.332 ⋄127.7%	1.431 ⋄160.2%	9.491 ⋄70.3%	0.224 ⋄92.0%	0.960 ⋄139.9%	53.31
DLabV3+ [8]	♢	1.347 ⋄28.4%	0.683 ⋄10.7%	3.256 ⋄14.3%	5.632 ⋄49.9%	0.374 ⋄15.0%	1.993 ⋄15.6%	40.35
SegFormer [59]		1.198 ⋄19.4%	0.741 ⋄2.0%	3.543 ⋄5.1%	5.957 ⋄52.4%	0.401 ⋄3.6%	2.167 ⋄0.2%	27.25
HRNet [52]		14.897 ⋄93.5%	0.196 ⋄285.7%	1.859 ⋄100.3%	17.984 ⋄84.3%	0.045 ⋄855.6%	0.525 ⋄338.7%	28.54
ViTPose [61]	♣	4.303 ⋄77.6%	0.376 ⋄101.1%	1.456 ⋄156.3%	5.545 ⋄49.1%	0.246 ⋄74.8%	0.805 ⋄186.1%	89.99
HRFormer [66]		1.259 ⋄23.4%	0.729 ⋄3.7%	3.479 ⋄7.0%	5.855 ⋄51.8%	0.393 ⋄9.4%	2.109 ⋄9.2%	10.10
HSNet [33]	♠	2.014 ⋄52.1%	0.431 ⋄75.4%	1.922 ⋄93.7%	3.016 ⋄6.4%	0.234 ⋄83.8%	1.007 ⋄128.7%	26.13
Ours		**0.965**	**0.756**	**3.732**	**2.823**	**0.430**	**2.303**	36.32

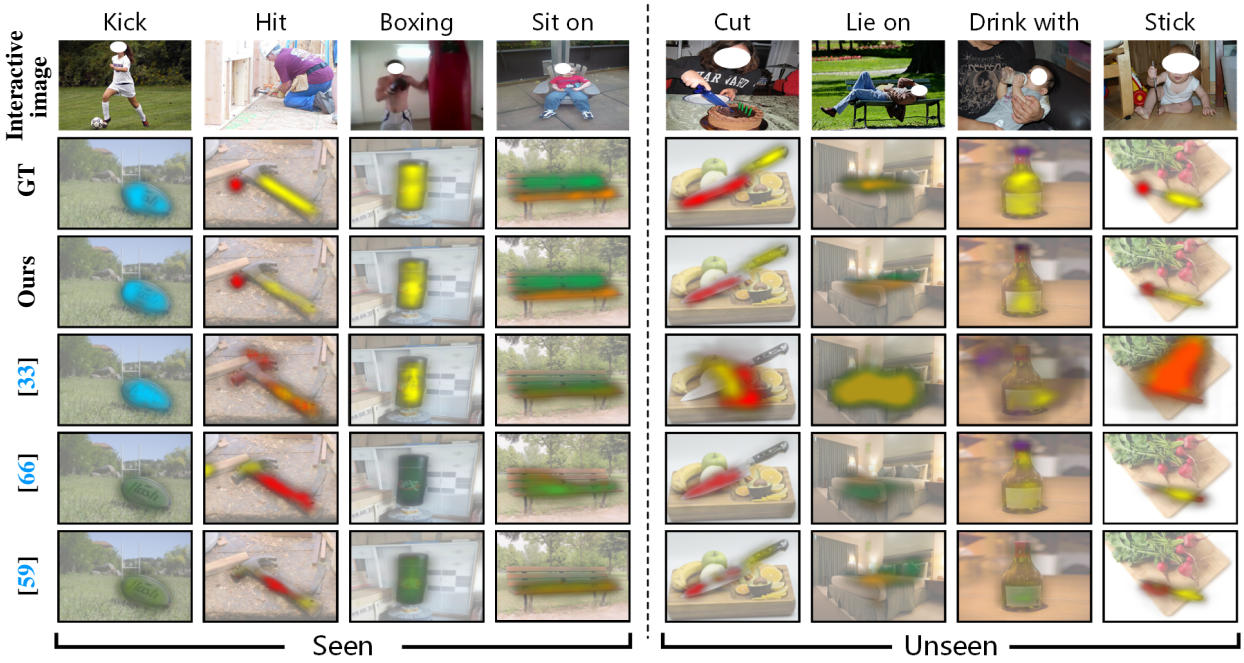

Figure 5. **Visualization of prediction results.** We show the visualization results of our model, few-shot segmentation (HSNet [33]), the best human pose estimation model (HRFormer [66]) and the segmentation model (SegFormer [59]).

forms the segmentation, human pose estimation, and few-shot segmentation models in all metrics at both the **Seen** and **Unseen** settings. Taking KLD as the metric, our method exceeds the best segmentation method by **19.4**%, surpasses the best human pose estimation model by **23.4**%, and outperforms the few-shot model by **52.1**% at the **Seen** setting. At the **Unseen** setting, our approach outperforms the best semantic segmentation method by **52.4**%, the best human pose estimation model by **51.8**%, and the few-shot segmentation model by **6.4**%. Moreover, our method significantly exceeds HSNet [33] in both SIM and NSS metrics. The performance advantage of our model is even more pronounced at the **Unseen** setting. It validates that exploiting the intrinsic connection between the body part and the object's local interactions using interactive affinity helps our model better generalize to unseen interactions and objects.

Fig. 5 shows the visualization of affordance maps. Our method outperforms representative models at both **Seen** and **Unseen** settings. For the rugby ball and the punching bag, it is difficult to identify the corresponding interactions using the segmentation methods due to the multiple possibilities of affordance. It suggests that the interactive affinity can provide valuable hints about the interconnectedness of the human body part with the local region of the object, thereby helping reduce the ambiguity of the perceived action possibilities. For knives to be used for both "Cut" and "Stick", our model still recognizes the difference between interactions ("Outside"). It suggests that using pose

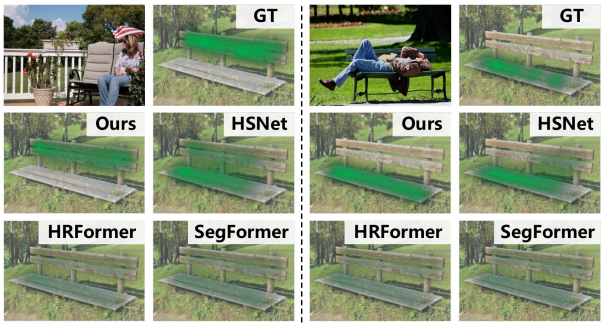

Figure 6. **Different interactive images** *w.r.t* **Same non-interactive images.** We show the results for predicting the back contact region under the interaction of "Sit on" and "Lie on".

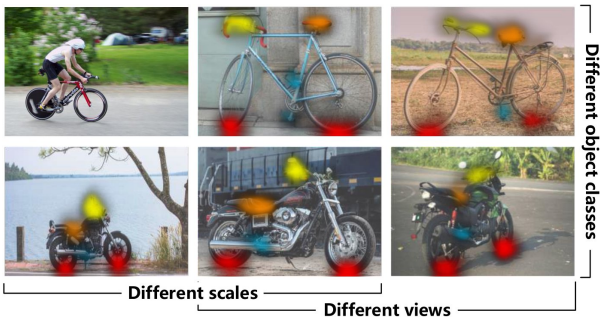

Figure 7. **Different interactive images** *w.r.t* **Same non-interactive images.** We present the prediction results of the model for different non-interactive images of the same interactive image.

data as guidance can assist the network in identifying the region where the object is in contact with the outside world to complete the interaction. Fig. 6 shows the results of different interaction images for the same non-interactive image. Since "Sit on" and "Lie on" correspond to different contact regions on the back, we only visualize this part. Our model can perceive the variations between interactions and accurately transfer the interaction-related invariant features to the corresponding regions. Fig. 7 shows the prediction for the same interactive image corresponding to different non-interactive images. Although different object categories, various scales, and different views, our model can transfer the interactive affinity correctly for each part, which indicates that the IFE module can effectively establish the connection between the interactable regions from different branches and facilitate the transfer of interactive affinity.

4.4. Ablation Study

The ablation study results are shown in Table 2. Our framework shows a more pronounced improvement at the **Unseen** setting, suggesting that our method can generalize and transfer even for unseen interactions and objects with only a few interactive images. In particular, the KHP scheme shows a more evident improvement in the model's

Table 2. **Ablation study.** We investigate the influence of the IFE module and KHP scheme on model performance.

IFE	KHP	Seen			Unseen		
		KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
		1.208	0.728	3.491	5.593	0.396	2.170
✓		1.126	0.733	3.625	4.558	0.404	2.143
	✓	1.164	0.746	3.667	3.576	0.424	2.257
✓	✓	**0.965**	**0.756**	**3.732**	**2.823**	**0.430**	**2.303**

Table 3. **Ablation study of the pose guidance.** We investigate the impact of different pose feature fusion methods in the KHP scheme. "GAP" means global average pooled pose node features. "Conv" means concatenation with the contact feature and then being fused by a convolutional layer.

Mode	Seen			Unseen		
	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
GAP+Conv	1.317	0.715	3.517	4.053	0.383	1.962
Part+Conv	1.169	0.733	3.612	3.933	0.403	2.215
GAP+CT	1.253	0.710	3.578	4.223	0.386	1.937
Part+CT	**0.965**	**0.756**	**3.732**	**2.823**	**0.430**	**2.303**

performance. It indicates that obtaining interactive affinity from human-object interaction can better exploit the connections between the body part and the local region of the object, thus suppressing the multiple possibilities of interaction diversity. Besides, we investigate the effect of the pose guidance on the results. As shown in Table 3, "GAP" loses the association between different part keypoints and performs worse than using the keypoints corresponding to different parts separately. "Conv" fusion is worse than the cross transformer, possibly due to the fact that the transformer explicitly computes the correlation between the contact region and the body keypoints, thus enabling better mine the interactive affinity representation.

4.5. Performance Analysis

Different Classes. Fig. 8 shows the results of different models on each affordance category. Although our model does not outperform HSNet in some categories [33] at the **Unseen** setting (the scores are comparable), it still outperforms other methods, which indicates that our method has a good generalization ability when facing unseen objects. For the overlap of objects in "Cut", "Stick" and "Lift", a finer-grained interactive affinity representation can suppress the effects of multiple possibilities and achieve better results.

Different body parts. Table 4 shows the results regarding different body parts. Our model achieves almost the best results at the **Seen** setting. At the **Unseen** setting, it also achieves excellent transfer performance in most interaction regions corresponding to the body parts. As most of the interaction is done through hand-object contact, the results of our model at the **Unseen** setting are much better

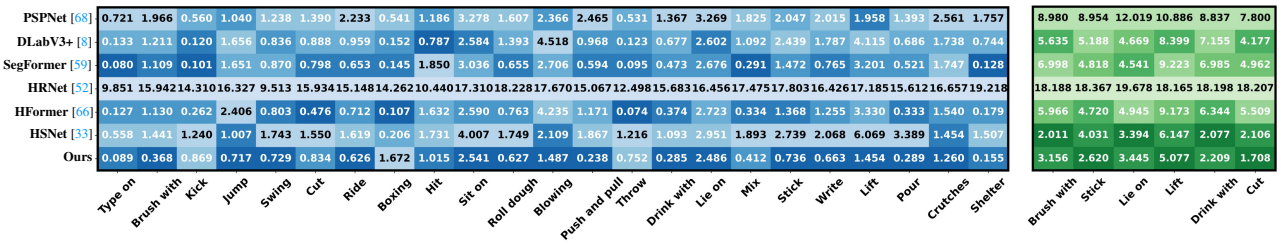

Figure 8. **Different classes.** We measure the KLD metric for each affordance category, with darker colors representing higher performance. The left and right represent the results at the **Seen** and **Unseen** settings, respectively.

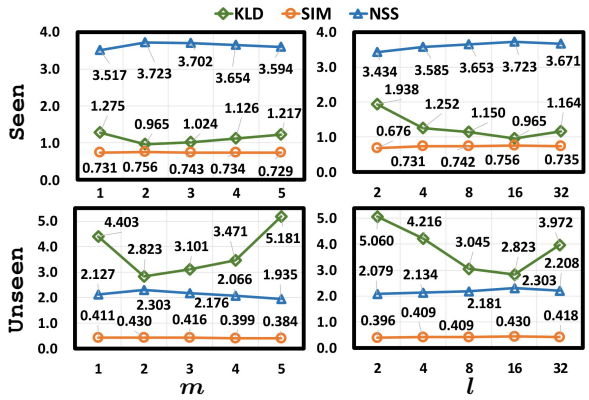

Figure 9. **Different Hyper-parameters.** We investigate the impact of the number of layers m and tokens l in the interactive enhancement module on the experimental results.

than all the other methods. It demonstrates that our method uses interactive affinity to explore the interconnections between body parts and local regions of objects, thus reducing the ambiguity in perceiving action possibilities and enabling accurate affordance knowledge transfer.

Different hyper-parameters. Fig. 9 shows the impact of the hyper-parameters l and m in the IFE module on the model's performance. The influence at the **Unseen** setting is slightly more apparent, indicating that these parameters are more sensitive to the generalization and transfer of unseen objects. The number of cross-branch tokens greatly affects the model performance. Too small l may lead to difficulties in fully expressing the contextual cues of the different branch interactable features, with the best results obtained at $l = 16$. The model gives the best results with the number of layers $m = 2$. The smaller m makes it difficult to fully aggregate the contextual cues and establish associations in the interactable regions. As m increases, it increases the parameters of the model and the complexity of the optimization, leading to negative results.

Limitations. As shown in Table 4, our method still has a large gap between "Back" and the other body parts. It is mainly due to the sample imbalance as fewer examples contain "Back". Besides, the sitting posture occludes most of the region where the object interacts with the back, resulting in an incomplete back contact region feature extraction. In future work, we will consider addressing the data imbalance [16, 25, 41] issue and exploiting the co-relation relationship between body parts.

Table 4. **Different parts.** We evaluate the predictions for each body part and object contact region. The **Bold** and **Underline** represent the best and sub-optimal results, respectively.

	Method	Hand	Feet	Mouth	Hips	Back	Outside
Seen	PSPNet	1.471	1.284	1.860	1.483	7.818	1.337
	DLabV3+	0.775	1.195	3.069	0.576	7.165	0.945
	SegFormer	**0.547**	1.293	2.142	0.606	7.587	0.847
	HRNet	14.583	17.202	17.095	14.197	24.823	13.208
	HFormer	0.634	1.741	2.573	0.430	7.152	0.844
	HSNet	1.640	1.518	4.561	1.283	9.141	1.218
	Ours	0.687	**1.151**	**0.816**	**0.311**	**6.975**	**0.570**
Unseen	PSPNet	6.040	-	15.187	11.321	19.479	8.043
	DLabV3+	4.393	-	11.308	1.366	11.397	4.137
	SegFormer	3.989	-	13.015	1.310	11.184	4.673
	HRNet	17.761	-	19.257	17.221	25.737	17.830
	HFormer	4.477	-	11.551	1.439	12.196	4.393
	HSNet	2.100	-	**5.259**	1.099	**7.984**	2.356
	Ours	**1.612**	-	6.084	**0.803**	8.737	**1.855**

5. Conclusion

This paper proposes to leverage interactive affinity for effective affordance learning by counteracting the influence of multiple possibilities. To this end, a pose-aided interactive affinity learning framework is introduced, which is able to exploit pose data to guide the network to mine the interactive affinity representation of body parts and object local contact from human-object interaction. Furthermore, we constructed a contact-driven affordance learning (CAL) dataset by collecting and labeling over $5,000$ images from 23 affordance categories. Our model outperforms six representative models in three related fields and can serve as a strong baseline for future affordance learning research.

Acknowledgments. This work is supported by National Key R&D Program of China under Grant 2020AAA0105701 and ARC FL170100117 and IH180100002.

References

[1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 4

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3

[3] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017. 1

[4] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 5

[5] 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. 5

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 5

[7] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015. 1

[8] 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. 5, 6, 8

[9] 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. 1, 2

[10] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2021. 2, 3

[11] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 1, 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Vibekananda Dutta and Teresa Zielinska. Action prediction based on physically grounded object affordances in human-object interactions. In *2017 11th International Workshop on Robot Motion and Control (RoMoCo)*, pages 47–52. IEEE, 2017. 1

[14] 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. 2, 3, 5

[15] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977. 1

[16] Khan Md Hasib, Md Iqbal, Faisal Muhammad Shah, Jubayer Al Mahmud, Mahmudul Hasan Popel, Md Showrov, Imran Hossain, Shakil Ahmed, Obaidur Rahman, et al. A survey of methods for managing the classification and solution of data imbalance problem. *arXiv preprint arXiv:2012.11870*, 2020. 8

[17] 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. 3

[18] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 1, 3

[19] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration*, pages 181–184. Citeseer, 2011. 2

[20] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, 2016. 3

[21] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 5

[22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 2021. 3

[23] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[24] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[25] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018. 8

[26] Yi-Chi Liao, Kashyap Todi, Aditya Acharya, Antti Keurulainen, Andrew Howes, and Antti Oulasvirta. Rediscovering

affordance: A reinforcement learning perspective. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022. 3

[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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. *arXiv preprint arXiv:2106.14747*, 2021. 2, 3, 5

[30] 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. 2, 3, 5

[31] Wei Mao, Miaomiao Liu, Richard Hartley, and Mathieu Salzmann. Contact-aware human motion forecasting. *arXiv preprint arXiv:2210.03954*, 2022. 3

[32] Huaqing Min, Ronghua Luo, Jinhui Zhu, Sheng Bi, et al. Affordance research in developmental robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):237–255, 2016. 3

[33] 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. 5, 6, 7, 8

[34] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 3

[35] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning*, pages 1666–1677. PMLR, 2022. 3

[36] Roozbeh Mottaghi, Connor Schenck, Dieter Fox, and Ali Farhadi. See the glass half full: Reasoning about liquid containers, their volume and content. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1871–1880, 2017. 2

[37] 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. 1, 3

[38] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1, 2, 3

[39] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. *Advances in Neural Information Processing Systems*, 33:2005–2015, 2020. 1, 3

[40] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770. IEEE, 2016. 1, 3

[41] Minlong Peng, Qi Zhang, Xiaoyu Xing, Tao Gui, Xuanjing Huang, Yu-Gang Jiang, Keyu Ding, and Zhigang Chen. Trainable undersampling for class-imbalance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4707–4714, 2019. 8

[42] 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. 5

[43] Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu. Predicting human activities using stochastic grammar. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1164–1172, 2017. 1

[44] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *International Journal of Computer Vision*, 129(5):1616–1649, 2021. 1

[45] 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. 3

[46] Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1383–1391, 2017. 1, 3

[47] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804, 2017. 3

[48] Giulio Schiavi, Paula Wulkop, Giuseppe Rizzi, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Learning agent-aware affordances for closed-loop interaction with articulated objects. *arXiv preprint arXiv:2209.05802*, 2022. 1, 3

[49] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision*, pages 516–533. Springer, 2022. 3

[50] 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. 1

[51] Tianmin Shu, Xiaofeng Gao, Michael S Ryoo, and Song-Chun Zhu. Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1669–1676. IEEE, 2017. 1

[52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 5, 6, 8

[53] Yuan Sun, Shuyue Fang, and Zuopeng Justin Zhang. Impression management strategies on enterprise social media platforms: An affordance perspective. *International Journal of Information Management*, 60:102359, 2021. 1

[54] Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision (IJCV)*, 7(1):11–32, 1991. 5

[55] Karthik Mahesh Varadarajan and Markus Vincze. Parallel deep learning with suggestive activation for object category recognition. In *International Conference on Computer Vision Systems*, pages 354–363. Springer, 2013. 3

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[57] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2596–2605, 2017. 2

[58] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European Conference on Computer Vision*, pages 90–107. Springer, 2022. 2

[59] 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, 4, 5, 6, 8

[60] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. *arXiv preprint arXiv:2204.02445*, 2022. 3

[61] 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. 5, 6

[62] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. ViTAE: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[63] Natsuki Yamanobe, Weiwei Wan, Ixchel G Ramirez-Alpizar, Damien Petit, Tokuo Tsuji, Shuichi Akizuki, Manabu Hashimoto, Kazuyuki Nagata, and Kensuke Harada. A brief review of affordance in robotic manipulation research. *Advanced Robotics*, 31(19-20):1086–1101, 2017. 1, 3

[64] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 4

[65] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 3

[66] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. 2021. 5, 6, 8

[67] 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. 2, 3, 5

[68] 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. 5, 6, 8

[69] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015. 1