Visual-geometric Collaborative Guidance for Affordance Learning

Hongchen Luo[1,2], Wei Zhai[1], Jiao Wang[2], Yang Cao[1] and Zheng-Jun Zha[1]

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. To this end, we propose a visual-geometric collaborative guided affordance learning network that incorporates visual and geometric cues to excavate interactive affinity from human-object interactions jointly. Particularly, a semantic-pose heuristic perception (SHP) module is devised to exploit both semantic and geometric cues to guide the network to focus on interaction-relevant regions, alleviating the effects of combinatorial relational ambiguity. Meanwhile, A geometric-apparent alignment transfer module is introduced to jointly align local regions of apparent and structural similarity, eliminating the transport difficulties posed by intra-class correspondence ambiguity. Besides, a contact-driven affordance learning (CAL) dataset is constructed by collecting and labeling over 55,047 images. Experimental results demonstrate that our method outperforms the representative models regarding objective metrics and visual quality. Project: github.com/lhc1224/VCR-Net.

Index Terms—Embodied AI, Affordance Learning, Interactive Affinity, Benchmark

I. INTRODUCTION

THE objective of affordance learning is to locate the "action possibilities" regions of an object [2], which is crucial in the embodied intelligence field. For an intelligent agent in an interacting environment, 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, 4]. Moreover, affordance learning has a wide range of applications in fields such as action recognition [5], scene understanding
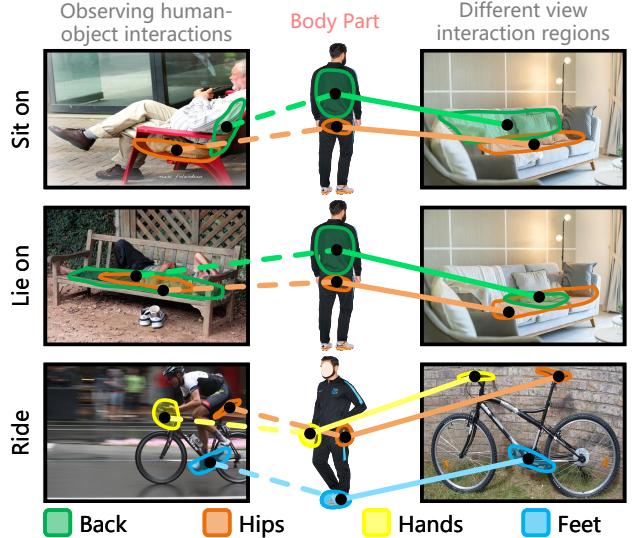
[1] University of Science and Technology of China, Hefei, China
[2] Northeastern University, Shenyang, China

Hongchen Luo (luohongchen@ise.neu.edu.cn), Wei Zhai (wzhai056@ustc.edu.cn), Jiao Wang (wangjiao@ise.neu.edu.cn), Yang Cao (forrest@ustc.edu.cn), Zheng-Jun Zha (zhazj@ustc.edu.cn)

A preliminary version of this work has appeared in CVPR 2023 [1].

Fig. 1. **Interactive affinity.** **(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.

[6], human-robot interaction [7], autonomous driving [8] and VR/AR [9].

Affordance is a dynamic property closely related to humans and the environment [10]. Previous works [11, 12, 13, 14] 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 [4, 15, 16] 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 unseen scenarios [17]. To this end, researchers consider learning from human demonstration in an action-free manner [3, 18, 19, 20, 21]. 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 (as shown 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 (as shown in Fig. 1 (b)).

However, multiple ambiguities in the interaction process render it challenging for the model to perceive and robustly generalize the interactive affinity representation. This is mainly reflected in the ambiguity of combinatorial relations and intra-class correspondence. The combinatorial relationship ambiguity means that due to the diversity of human-object interactions, the combination of interactions between the body and the object's local regions is complex and various, resulting in the model's difficulty in perceiving the contact regions corresponding to distinct interactions and accurately mining the interactive affinity representations. To address this problem, we consider leveraging semantic and geometric cues to guide the model in mining interactive affinity (Fig. 2 (a)). We leverage the textual semantics to explicitly align visual features to guide the model to focus on interaction-related contact regions while exploiting the geometric relationships between various body joints to mine potential structural cues in the interactable regions of an object, enabling the model to focus on complex interactive combinatorial relationships adequately. The intra-class correspondence ambiguity refers to the fact that since the same affordance covers multiple classes of objects, there are significant variations in the appearance, views, and scales, which makes the representations of the interactable local regions corresponding to the objects and the relative spatial relationships in the interactive and non-interactive images more inconsistent, resulting in the possible occurrence of negative transfer. To this end, we consider both apparent and geometric alignment to achieve accurate transfer (Fig. 2 (b)). We use dense matching to align visual similarity features between local interactable regions while utilizing the geometric structure of the human pose as a bridge to align the similar geometric structure of interactable regions, eliminating the effects of intra-class corresponding ambiguity by exploiting the complementary relationships.

In this paper, we propose a **V**isual-geometric **C**ollaborative guided affo**R**dance learning **Net**work (**VCR-Net**) for extracting interactive affinity representations in interactive images and transferring them to non-interactive images to achieve understanding and generalization of various complex interactions. Firstly, a **S**emantic-pose **H**euristic **P**erception (**SHP**) module is designed to guide the model to focus on interaction-relevant local contact regions, obtaining interactive affinity representations from diverse interactions. Subsequently, a **G**eometric-apparent **A**lignment **T**ransfer (**GAT**) module is introduced to transfer the interactive affinity to the non-interactive image. Specifically, the SHP module utilizes a cross-transformer to inject the semantic cues from different body parts into the interactive image's features. Meanwhile, it employs the deep equilibrium model [22] to adaptively adjust the geometrical correspondences between the image and mesh features, enabling the model to adapt to various complex postures. On the other hand, the GAT module adaptively adjusts the association between mesh and non-interactive image features using the deep equilibrium model
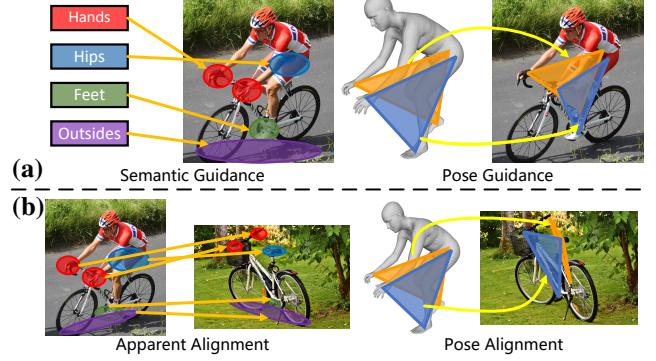

Fig. 2. **Motivation.** (a) We consider both the semantic and structural cues to extract the interactive affinity from the interaction images. (b) We exploit the implicit structural cues of body pose and apparent similarity to jointly perform the interactive affinity transfer.

during the interactive affinity representation transfer process. and then it considers the contact region features extracted from the interactive image and the non-interactive features to calculate the feature similarity between pixels to activate the corresponding interactable local regions.

Although the numerous related datasets [23, 19, 24, 6, 25, 21] 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 research, this paper constructs an **C**ontact-driven **A**ffordance **L**earning (**CAL**) dataset, consisting of $55,047$ images from 35 affordance and 61 object categories. We conduct contrastive studies on the CAL dataset against eight representative models in several related fields. Experimental results validate the effectiveness of our method in solving the multiple possibilities of affordance. In summary, our primary contributions are:

1) This paper considers 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) A novel visual-geometric collaborative guided affordance learning network is proposed that utilizes visual and geometric cues to achieve more accurate interactive affinity representations extracted from human-object interactions, eliminating the effects of multiple ambiguities.

3) Experiments on the CAL dataset demonstrate that our VCR-Net outperforms state-of-the-art methods and can serve as a strong baseline for affordance learning research.

This paper builds upon our conference version [1], which has been extended in three distinct aspects. **Firstly**, we provide an in-depth analysis of the interactive affinity extraction and the transfer process. **Secondly**, we introduce a semantic-pose heuristic perception module that leverages semantic and geometric structures jointly to guide the model to extract the interactive affinity representation more accurately. **Thirdly**, we introduce a geometric-apparent alignment transfer module to eliminate inconsistencies during transfer by aligning structural and apparent similarities. **Fourthly**, we expand the dataset nearly ten times in several aspects and conduct more experiments to analyze the model's performance comprehensively.

The remainder of this paper is organized as follows: Sec.

II provides a brief review of existing related studies. Sec. III describes the pipeline of the proposed model and its details. In Sec. IV, we introduce the collection, annotation process, and statistical analysis of the CAL dataset. Sec. V describes the experimental setting and provides comprehensive results and analysis. In Sec. VI, we present the conclusions, limitations, and future directions of this work.

II. RELATED WORK

A. Affordance Learning

In recent years, the rapid growth of robotics, autonomous driving, and embodied intelligence has thrust affordance learning into prominence [10]. Early works [11, 12, 14, 26, 27] mainly adopt the label assignment paradigm, positing an intrinsic correlation between functional regions and their corresponding affordance labels. However, it encounters challenges in addressing the multiplicity of affordance interpretations, which arise from variations in environmental contexts and operator behavior. Recent studies have integrated reinforcement learning, leveraging tailored reward functions to enable agents to dynamically interact with their environment and refine their affordance perception skills [4, 28, 15]. This series of works can mitigate the problem of generalization to diverse environments. However, the actions and states are in a high-dimensional and complex space, which is difficult to optimize using autonomous exploration and is restricted by the scenarios of the simulation platform, which prevents it from being applied in various complex scenarios and tasks. Some studies [29, 30, 31] achieve generalization of articulated object manipulation by decoupling the object part while aligning the relationship between semantics and interactable parts. While some other works consider learning the object's affordance from the human demonstration in an action-free manner [20, 18, 19, 23, 3, 32, 33, 34], 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. Then, Yang et al. [21] proposed a multi-task learning framework named GAAF-Dex that learns grasping knowledge from the exocentric view and transfers it to the egocentric view, further advancing interactive skill learning. 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 yet to resolve the multiple possibilities issue fully. In contrast, this paper considers using the inherent cues of interconnectivity between humans and objects to reduce the ambiguity of the perceived action possibilities.

B. Interaction Relationship Learning

Since studying the human-object interaction relationship is crucial for improving several areas, such as the autonomy of intelligent agents and the interaction of environments, this has attracted a wide range of attention from researchers. Some work [35, 36, 37] mainly considered detecting human-object interaction pairs from images. However, the diversity of interactions and the variability in actions for different objects complicate the mapping between interaction labels and visual cues, posing challenges for accurate perception and generalization. Some efforts consider contact relationships between human beings and objects as cues to achieve a more accurate understanding of interactions. Shimada et al. [38] use body-scene contacts to guide 3D human capture. Bhatnagar et al. [39] 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. [40] introduce distance-based contact maps as an explicit constraint for human motion forecasting. Yang et al. [41] propose to model hand-object interaction by explicitly representing the contact using the Contact Potential Field (CPF). Similar to the above work, we utilize the interactive affinity to mine the interconnectivity between humans and objects, helping reduce the ambiguity in affordance learning.

III. 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 **VCR-Net** is shown in Fig. 3. It first extracts features through a transformer [42] backbone to obtain $\mathbb{X}^{in} = \{\boldsymbol{X}_i^{in}, i \in [1,4]\}$ and $\mathbb{X}^{non} = \{\boldsymbol{X}_i^{non}, i \in [1,4]\}$, respectively ($i$ indexes the block of the backbone). For human poses, we use a pose encoder consisting of several layers of transformers [43] to obtain the feature representation $\boldsymbol{X}^P$. For different body part descriptions, the text features are extracted using Bert [44]: $\boldsymbol{X}^T = \text{Bert}(\boldsymbol{T})$. Subsequently, an semantic-pose heuristic perception (**SHP**) module collectively guides the network through semantics and pose to obtain affinity representations from human-object interaction images (Sec. III-A). Finally, the geometric-apparent alignment transfer (**GAT**) module guides the network to transfer the interactive affinity representation to the non-interactive image through the geometric prior of the pose and the local region's apparent similarity (Sec. III-B).

A. Semantic-pose Heuristic Perception

Due to the complexity and variety of human-object interaction relationships, there exists combinatorial ambiguity between human-object interaction contacts, rendering it difficult to perceive the interactive affinity accurately. However, the collaborative relationships between body parts are similar to human-object interactions. To this end, we consider using the rotational offset relationship of different joint locations in the 3D mesh [45] to direct the model's focus on the corresponding interactable regions. Meanwhile, to strengthen the module's ability to perceive different body parts, we introduce semantic guidance, which enhances the network's perception of the affinity representation of different parts with objects by aligning the association between semantic descriptions of different parts and visual representations of different parts in the feature space during the training process. As shown in Fig. 3, firstly, we exploit the cross-transformer [43] to align the connection between semantic and visual features of the
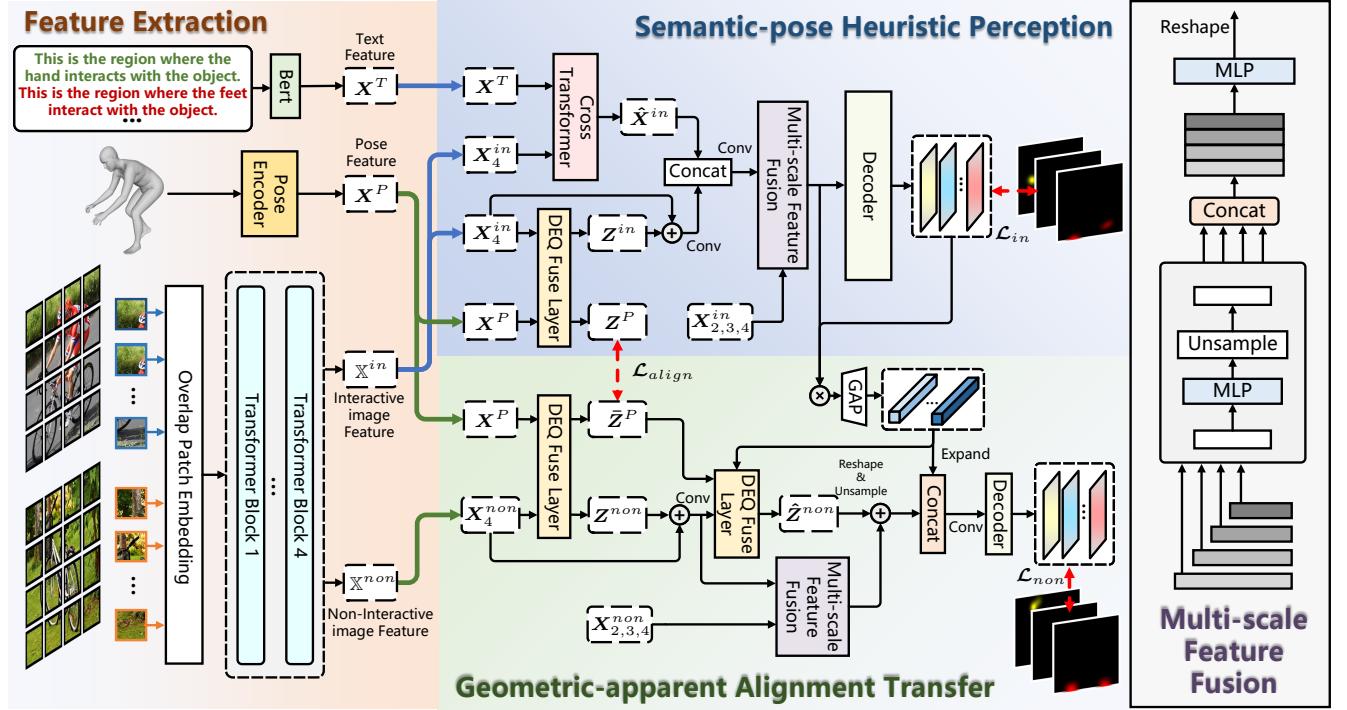
Fig. 3. **Overview of the proposed VCR-Net.** Our approach mainly consists of three parts: feature extraction, semantic-pose heuristic perception module (Sec. III-A) and geometric-apparent alignment transfer module (Sec. III-B).

interactive images which enhance the model's perception of the different body part contact regions. The cross transformer ($O = \text{CT}(X_1, X_2)$) is computed as:

$$Y = \text{MCA}(\text{LN}(X_1), \text{LN}(X_2)) + X_1, O = \text{MLP}(\text{LN}(Y)) + Y, \tag{1}$$

where X_1 represents query, X_2 represents key and value, $\text{MCA}()$ denotes the dot-production attention [46]. In the SHP module, we concatenate the features of text and interactive images as key and value, I^{in} as query fed into a one-layer cross transformer to guide the network to establish the connection between different body parts and visual features:

$$\hat{X}^{in} = \text{CT}(X^{in}, [X^{in}, X^T]). \tag{2}$$

where $[\cdot, \cdot]$ represents the concatenate operation. On the other hand, human joints can provide rich geometric cues for the model to mine associations between the interactable local regions of an object. However, due to the large modal differences between pose features and image features and the large differences in the interactive images in various views, it is arduous to learn the links between the two modalities with a fixed network layer. To this end, we adaptively adjust the correspondence between the joints and the local regions of the objects in the interactive images by introducing the Deep Equilibrium Model [22] to dynamically sense effective information and remove redundancy. In general, the DEQ model forward process can be written:

$$Z^{[i+1]} = f_\theta(Z^{[i]}; X), \quad i = 0, .., L - 1. \tag{3}$$

DEQ denotes an infinitely deep network with only one layer of f_θ that converges to the equilibrium state and can be implicitly backpropagated in one calculation. The results are independent of the chosen root-finding algorithm or the network structure

of f_θ, which provides more flexibility in the design of f_θ. In the SHP module, for a given input $X = \{X_i\}$, we define the following form of f with $i \in [1, 2]$ for example, Q is calculated as follows:

$$Q = [W_1^Q X_1, W_2^Q X_2] + [W_1^{Q'} Z_1, W_2^{Q'} Z_2], \tag{4}$$

K and V are calculated similarly to Q and f_θ is expressed as follows:

$$f_\theta(Z; X) = \text{FFN}(Z + \text{Attention}(Q, K, V)) + Z. \tag{5}$$

Specifically, in order to utilize pose to guide interactive and non-interactive images to establish connections between interactable regions, we employ the following format:

$$Z^{in*} = \text{DEQFuse}([X_4^{in}, X^P]). \tag{6}$$

where Z^{in*} can be split into Z^P and Z^{in}. We sum Z^{in} with the original X_4^{in} features to obtain an interactable region feature representation that fuses the pose cues. Then, it is concatenated with X_4^{in} and fed into a layer of convolution to obtain an affinity representation X_{sp}^{in} that fuses the pose geometry information and semantic cues:

$$\tilde{X}^{in} = \text{Conv}(Z^{in} + X_4^{in}), \quad X_{sp}^{in} = \text{Conv}([\tilde{X}^{in}, \hat{X}^{in}]). \tag{7}$$

After obtaining the active affinity representation, we send it to a multi-scale feature fusion layer to fuse with the shallow features extracted from the backbone to obtain the higher resolution features [42]:

$$F_i^{in} = \text{Upsample}(\text{MLP}(X_{sp}^{in})), \quad i = 4, \tag{8}$$

$$F_i^{in} = \text{Upsample}(\text{MLP}(X_i^{in})), \quad i \in [1, 3], \tag{9}$$

$$\hat{F}^{in} = \text{MLP}([F_1^{in}, F_2^{in}, F_3^{in}, F_4^{in}]). \tag{10}$$
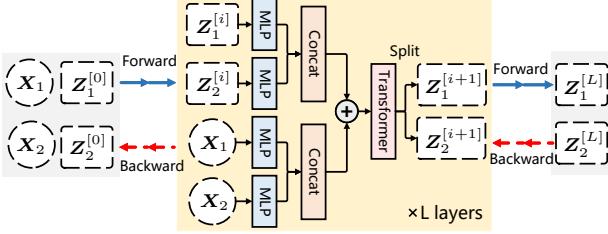
Fig. 4. **DEQ fuse layer.** The DEQ fuse layer consists of a transformation f_θ that is driven to equilibrium between different input features.

After fusing multi-scale features, $\hat{F}^{in}$ is reshaped to 2D feature map and sent to a decoder $\mathrm{Decoder}(\cdot)$ to obtain the contact prediction: $D^{in} = \mathrm{Decoder}(\hat{F}^{in})$. D^{in} contains N (the number of body parts) channels, which represents the regions of body parts interact with the object.

B. Geometric-apparent Alignment Transfer

Due to multiple classes of objects in the same affordance category, apparent, viewpoint, and scale variations create intra-class correspondence ambiguities, resulting in inaccurate transfer of the interactive affinity. To this end, we align structural and apparent aspects to achieve a more accurate and robust transfer. As shown in Fig.3, for geometric correspondences, we first use the pose as a bridge to activate regions in the non-interactive image that can match the geometric structures presented by the different joints of the body. Similar to the operations in the SHP module where pose guides the interactable region in the interactive image, we use the DEQ fusion layer to establish the link between the non-interactive image features and the pose features: $Z^{non*} = \mathrm{DEQFuse}([X_4^{non}, X^P])$. where Z^{non*} can be split into $\bar{Z}^P$ and Z^{non}: $\tilde{X}^{non} = \mathrm{Conv}(Z^{non} + X_4^{non})$. To bring the pose-guided interactions with the corresponding geometrical structures in the non-interactive images as close as possible, we introduce a pose alignment loss $\mathcal{L}_{align}$ that is used to narrow down the features of the poses in the two DEQ fusion layers:

$$\mathcal{L}_{align} = KLD(\bar{Z}^P, Z^P). \tag{11}$$

For the apparent matching relationship, we first input the pose, $\tilde{X}^{non}$, and G^{in} ($G^{in} = [G_1^{in}, ..., G_N^{in}]$, where $G_j^{in} = \mathrm{MASK}_i \otimes \hat{F}^{in}$) together into the DEQ fusion layer, which utilises the pose and the corresponding mask region in Z^{non} to orientate the network to focus on the local features in the contact region that are related to the pose:

$$\hat{Z}^{non*} = \mathrm{DEQFuse}([\tilde{X}^{non}, G^{in}, \bar{Z}^P]. \tag{12}$$

Non-interactive image features $\hat{Z}^{non}$ can be obtained by $\hat{Z}^{non*}$ split. Meanwhile, $\tilde{X}^{non}$ take the feature representation of the non-interactive branch Z^{non} and compute the high-resolution feature output $\hat{F}^{non}$ by following Eq. 8 $\sim$ Eq. 10. Subsequently, the pooled G_j^{in} is concatenated with $\hat{F}^{non}$ and R^{non} and fed into the convolutional layer to activate the regions where the features are apparently similar and go through the decoder to obtain the final output:

$$F^{fuse} = \mathrm{Conv}([\hat{F}^{non} + \mathrm{Upsample}(\hat{Z}^{non}), \mathrm{Expand}(G^{in})]), \tag{13}$$

TABLE I
THE DIMENSIONS, DOMAINS OF DEFINITION, AND MEANINGS OF THE SYMBOLS USED IN THE PROPOSED APPROACH.

	Dimensions	Meanings
I^{in}/I^{non}	$3 \times 224 \times 224$	Input image
P	53×3	Human pose
X_i^{in}/X_i^{non}	$c_i \times h_i \times w_i$	Image feature
X^P	$c \times 53$	Pose feature
Z^P	$c \times 53$	Pose features after DEQ fusion layer
Z^{in}	$c \times h_4 w_4$	Interactive features after DEQ fusion layer
$\hat{F}^{in}$	$c \times h_1 \times w_1$	Interactive features after multiscale fusion
Z^{non}	$c \times h_4 w_4$	Non-interactive features after DEQ fusion layer
$\bar{Z}^P$	$c \times 53$	Pose features after DEQ fusion layer
G^{in}	$c \times h_1 \times w_1$	Interactive contact region features
$\hat{Z}^{non}$	$c \times h_4 w_4$	Non-interactive features after DEQ fusion layer
$\hat{F}^{non}$	$c \times h_1 \times w_1$	Non-interactive features after multiscale fusion
D^{in}/D^{non}	$N_{cls} \times h_1 \times w_1$	Contact region prediction

$$D^{non} = \mathrm{Decoder}(F^{fuse}). \tag{14}$$

During training, the total loss is: $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{in} + \lambda_2 \mathcal{L}_{non} + \lambda_3 \mathcal{L}_{align}$, where $\mathcal{L}_{in}$ and $\mathcal{L}_{non}$ are the binary cross-entropy loss for the two branches, respectively. λ_1, λ_2 and λ_3 are the loss weight parameters and are both set to 1.

IV. DATASET

This section introduces the collection, annotation, and attribute analysis of contact-driven affordance learning (CAL) datasets. In particular, section IV-A describes the sources, selection criteria, and filtering process. Section IV-B illustrates the annotation rules and processing. Section IV-C provides the properties of the CAL dataset and the results of the statistical analyses.

A. Dataset Collection

In embodied intelligence applications, agents might be in a variety of scenarios and interact in a variety of complex manners. To this end, we select 35 interaction categories that occur frequently in daily life and cover various indoor and outdoor scenes. To obtain rich human-object interaction images, we select from several datasets containing rich interactions COCO [47], HICO [36] and Visual Genome (VG) [48]. For HICO, we filter the images based on the interaction category. For COCO, we first filter the candidate images based on the category of the object and then select them based on whether or not they contain interactions. For the VG dataset, we choose candidate images based on the semantics of the objects and the relationship categories contained in the images. We select images containing more clearly defined pairs of human-object interaction relationships as part of the dataset. Further, to increase the data's diversity, we select some interactive images from PAD [23] and AGD20K [19] datasets for supplementation. Whereas for non-interactive object images, we select non-interactive objects that satisfy the semantic and affordance conditions from the scene-rich and diverse COCO [47], PAD [18, 23], AGD20K [24, 19], and VG [48] datasets, and we discriminate images from these datasets based on the object semantic categories and manually remove images with ambiguities. Examples of interactive and non-interactive images in the dataset are shown in Fig. 5.

Fig. 5. **Dataset image examples.** Some examples of images and annotations from the CAL dataset.
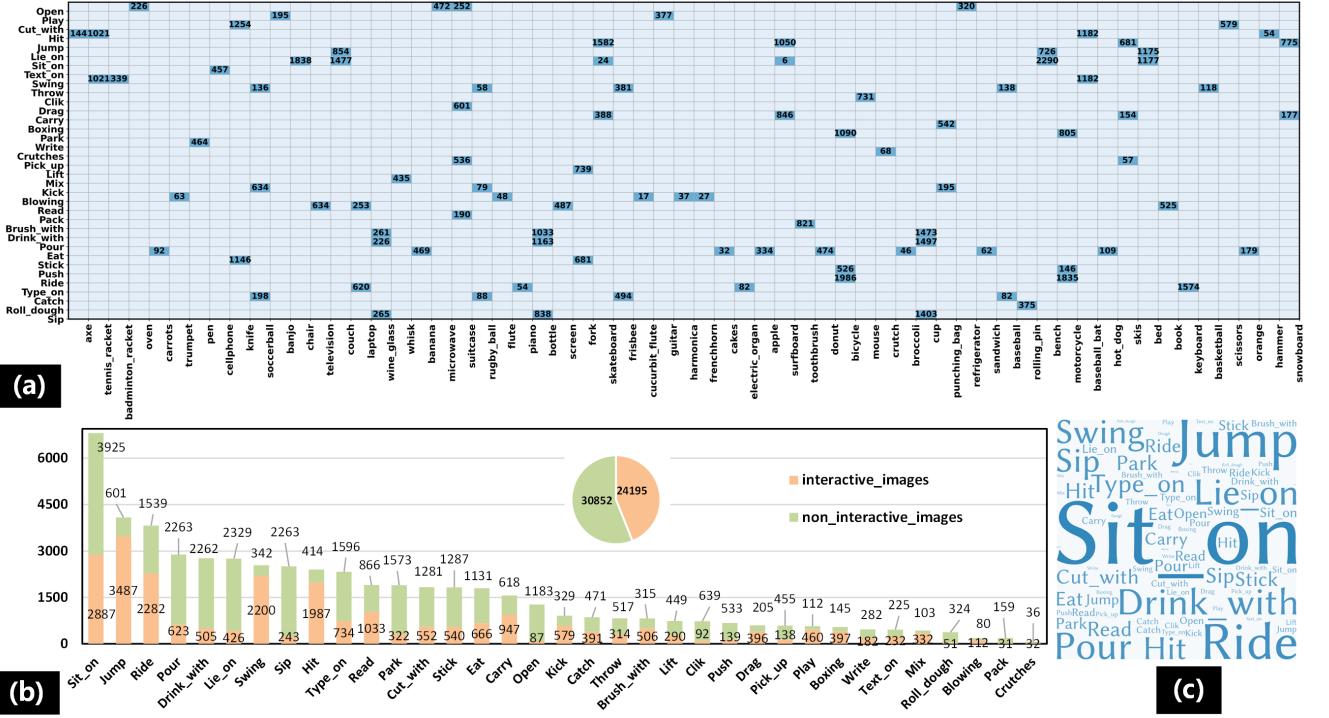

Fig. 6. **Some properties of Contact-driven Affordance Learning (CAL) dataset.** (a) Sample statistics of interactive and non-interactive images for each affordance category. (b) Affordance and object category confusion matrix, with numbers representing the samples of the current object category contained in the affordance category. (c) The word cloud distribution of affordances in the CAL dataset.

B. Dataset Annotation

We consider the interaction of 7 different body parts: "Hands", "Feet", "Mouth", "Hips", "Back", "Eye" 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 the object's "action possibilities" regions, the heatmap is appropriate for describing the possibility of interactions. We refer to the previous annotation works [20], [49], [50] and choose to annotate local regions of the image with different densities of points. We also assign object labels to the interactive and non-interactive images according to their object categories. Following Gebru et al. [51], we employ 10 random volunteers from the laboratory, ensuring that 3 volunteers annotated each image. For the interactive image annotation, we establish the following rules before labeling: **(1)** According to the given interaction category, select all the interaction instances in the image that satisfy the conditions, annotate humans and objects with the given interaction

bounding box, and simultaneously select the corresponding object semantic category. **(2)** For a given interaction instance, the hand ("point1"), feet ("point2"), mouth ("point3"), back ("point4"), hip ("point5"), eye ("point6"), and the corresponding region of the external interaction on the object ("point7") are labeled. The region of interaction with the outside refers to the local region of the object that affects the outside during the interaction between the human and the object, *e.g.*, the front half of a knife for cutting, a fork for spearing food. **(3)** Ignore the differences between the left & right in hands and feet. **(4)** The points are marked more intensively for regions with frequent interaction.

For the non-interactive image annotation, we establish the following rules before labeling: **(1)** Based on the examples in the interactive image, all the instances in the non-interactive image that satisfy the conditions are selected with labeling bounding boxes and corresponding object semantic categories. **(2)** For each instance of an image satisfying the conditions, the corresponding hand ("point1"), feet ("point2"), mouth ("point3"), back ("point4"), hip ("point5"), eye ("point6"), and the corresponding region of the outside interaction on the
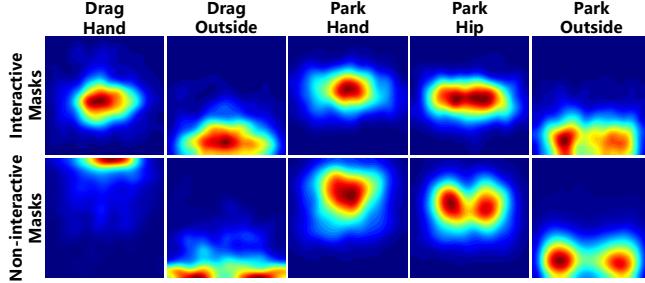
Fig. 7. **Affordance masks.** Average mask visualization of different body contact regions in the interactive and non-interactive images.

object ("point7") are annotated according to the examples in the interactive image. **(3)** For interactive local regions on the object, annotations of various densities are provided according to the frequency of interaction.

In processing the annotations, we set each annotation point on the mask to 1. Then, we utilize Gaussian blurring for each annotation point on the mask. The length and width of the image determines the size of the Gaussian kernel (ks):

$$ks = \frac{\sqrt{h^2 + w^2}}{\sigma(k)}, \quad where \quad \sigma(k) = 3, \tag{15}$$

then normalize it (*i.e.* $(V - V_{min})/(V_{max} - V_{min})$) to obtain the final mask. Some annotations are shown in Fig. 5.

C. Statistic Analysis

To get deeper insights into the CAL dataset, we show its important features from the following aspects. Fig. 6 (a) shows the number of images in each affordance category, which shows a long-tailed distribution of the data and an uneven distribution of samples among different categories. Most categories contain comparable quantities of interactive and non-interactive images, which provides a richer way of combining them for training and increases the diversity of the samples. Fig. 6 (b) shows the confusion matrix between different affordances and objects, which presents the property of multiple possibilities of affordances, *i.e.*, the same object may belong to more than one affordance category, and the property of multiple possibilities of one affordance category covering several different object categories. It increases the challenge of the task of affordance learning. Fig. 6 (c) exhibits the word cloud distribution of the affordance class of the CAL dataset, and it shows that we constructed the dataset to cover a wide range of affordance classes for various scenarios. We also visualized the affordance maps of different categories, as shown in Fig. 7. The masks of distinct affordance categories present diverse distribution regions, e.g., for interactive images, the location of the hand or hip contact region is mainly distributed in the middle of the image, which follows the recordings of the interaction content in daily life. On the other hand, non-interactive maps have different distribution regions from interactive maps. They are not always located in the center of the image, which makes it necessary for the model to reason about the corresponding affordance regions by using the relevant clues from the appearance and structure.

V. EXPERIMENTS

In this section, we first describe the experimental settings (Sec. V-A), which include the metrics, the comparison methods, and the implementation details. Then, we compare the performance of different methods in affordance learning, both subjectively and objectively (Sec. V-B). Following this, we analyze the performance of the different models from various perspectives (Sec. V-C). Finally, we conduct an ablation study of the modules in the methods to analyze the impact of different modules on the performance of affordance learning and generalization (Sec. V-D).

A. Experimental Settings

Metrics. Previous works mainly segment precise affordance regions [18, 6, 12], while the cross-view affordance grounding task considers a weakly supervised setting that predicts the affordance heatmap using only the affordance category label. Referring to the hotspots grounding-related works [20, 3], we adopt heatmaps to give a better description of the "action possibilities" (*i.e.*, affordance) and use **KLD** [58], **SIM** [59], and **NSS** [60] to evaluate the probability distribution correlation between the predicted affordance heatmap and Ground Truth (GT).

Comparison methods. In order to more thoroughly evaluate the advantages of our model on the affordance learning task, we compare representative models from multiple domains. Since this paper's task and segmentation model assign corresponding labels based on the image content, we select advanced models in the segmentation field for comparison (**PSPNet** [52], **DLabV3+** [53] and **SegFormer** [42]). Because our work needs to mine the interactive affinity from human-object interactions, we chose the classical human pose estimation models (**VitPose** [54] and **HRFormer** [55]) to evaluate the model's ability to perceive the interactive pose. On the other hand, the task setting of this paper is related to few-shot segmentation and text-guided image segmentation. Hence, we choose representative few-shot segmentation (**HSNet** [56]) and multimodal models (**CLIP** [57]) for comparison. Finally, we also compare the PIANet model [1] in the conference version to evaluate the performance of the affordance learning approach.

Implementation details. Our model is implemented in PyTorch and trained with the AdamW [61] optimizer. With random horizontal flipping, the input images are cropped to 224×224. We train the model for 35 epochs on a single NVIDIA RTX3090 GPU with an initial learning rate of 6e-5. We divide the dataset into **Seen**, **Obj Unseen**, and **Aff Unseen** settings, where the **Seen** setting splits the dataset into training, test, and validation sets according to 7:2:1. For the **Obj Unseen** setting, we select 54 categories of objects as the training set and the remaining 24 categories as the testing and validation set, followed by dividing this sample into the testing and validation sets according to 2:1, which is mainly used for evaluating the model's ability to understand and generalize the interactable regions of unseen objects. For the **Aff Unseen** setting, we select 24 of the affordance categories as the training set. The remaining portion

TABLE II

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, OBJ UNSEEN DENOTES THAT THE OBJECTS IN THE TRAINING AND TEST SETS DO NOT OVERLAP, AND AFF UNSEEN REPRESENTS THAT THE AFFORDANCE/OBJECTS IN THE TRAINING AND TEST SETS DO NOT OVERLAP.

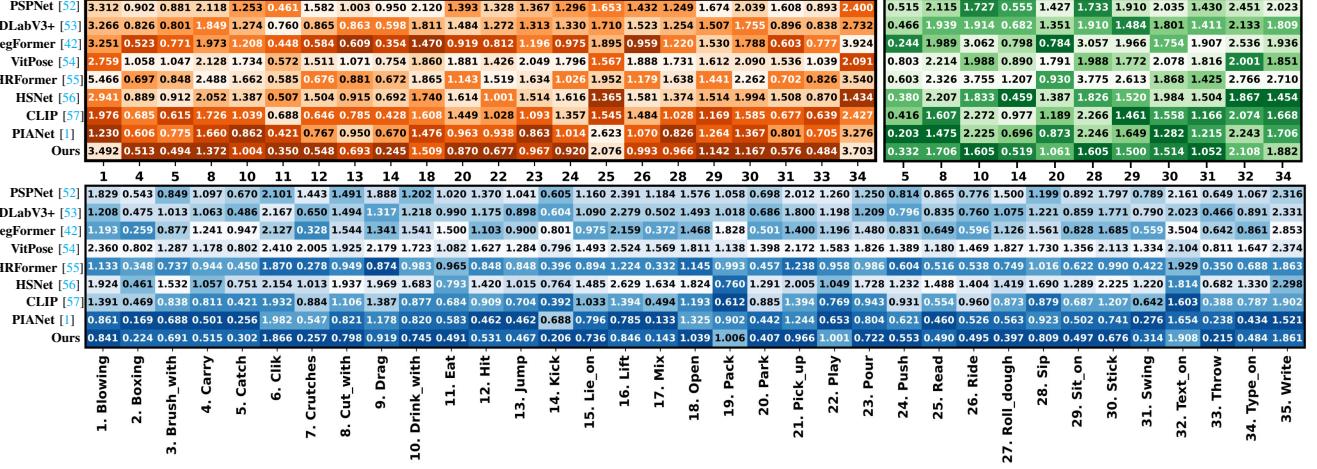| Method	Seen			Obj Unseen			Aff Unseen			params
	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	(M)
PSPNet [52]	1.095	0.493	2.362	1.547	0.388	1.624	1.779	0.316	1.175	53.32
DLabV3+ [53]	1.034	0.533	2.605	1.409	0.434	1.932	1.636	0.321	1.291	40.35
SegFormer [42]	1.060	0.648	3.211	1.240	0.539	2.533	2.023	0.363	1.486	27.35
ViTPose [54]	1.535	0.360	1.594	1.715	0.330	1.239	1.864	0.262	0.831	90.00
HRFormer [55]	0.778	0.616	3.106	1.443	0.511	2.353	2.434	0.352	1.403	10.11
HSNet [56]	1.511	0.341	1.426	1.663	0.310	1.215	1.774	0.279	1.130	28.13
CLIP [57]	0.883	0.538	2.944	1.252	0.456	2.236	1.605	0.342	1.534	89.88
PIANet [1]	0.630	0.680	3.444	1.194	0.541	2.616	1.597	0.362	1.601	36.32
Ours	**0.595**	**0.692**	**3.556**	**1.056**	**0.555**	**2.756**	**1.489**	**0.382**	**1.711**	39.69

Fig. 8. **Different Classes.** We measure the KLD, SIM, and NSS metrics for each affordance category, with darker colors representing higher performance. The blue, orange and green colours represent Seen, Obj Unseen and Aff Unseen, respectively.

is used as the testing and validation, dividing the testing and validation sets according to 2:1, mainly employed to assess the model's generalization ability for unseen objects with unseen interactions. In Seen and Obj Unseen settings, we set the batch size to 24 and train for 120,000 iterations, whereas in Aff Unseen, we take the batch size to 12 and train for 20,000 iterations.

B. Quantitative and Qualitative Comparisons

The experimental results are shown in Table II. Our approach achieves the best performance on multiple metrics across all settings, taking the KLD metrics as an example, with the Seen setting, our method improves **43.87%** compared to the best segmentation model, increases **23.52%** compared to the best human pose estimation method, gains **60.62%** relatively compared to the few-shot segmentation approach, exceeds the multimodal model by **32.61%**, and improves **5.56%** compared to the affordance learning's model. In the Obj Unseen setting, our model outperforms the advanced segmentation model by **14.84%**, superior to the best human pose estimation method by **26.82%**, higher than the network of few-shots by **36.50%**, with a relative improvement of **15.66%** compared to the multimodal model, and outperforms PIANet by **11.56%**. In the Aff Unseen setting, our model exceeds the advanced segmentation model by **26.40%**, excels the best human pose estimation method by **38.83%**, outperforms the

few-shot segmentation network by **16.07%**, and has a relative improvement of **7.23%** compared to the multimodal model, and outperforms PIANet by **6.76%**. Meanwhile, we visualize the affordance prediction results of the different methods, as shown in Fig. 9. Our method accurately predicts the affordance region of an object in all three settings, Seen, Obj Unseen, and Aff Unseen. Specifically, our approach works better to locate interactive object localised regions in non-interactive images without activating other irrelevant object regions in the interactive image scene (*e.g.*, the fourth row). For cases where the scene contains multiple objects (*e.g.*, the second, third, sixth, and tenth rows), the model is also capable of positioning all of them, suggesting that our consideration of pose as a cue between interactive and non-interactive images can help the model to more accurately perceive candidate regions based on the structure of the objects. For cases where the same object may belong to more than one affordance category (*e.g.*, rows 7 and 8), our method can also predict the corresponding affordance maps based on different body-contact regions, eliminating the ambiguity caused by the multiplicity of possibilities of affordance.

C. Performance Analysis

Different classes. To evaluate the perceived ability in different categories, we show the KLD metrics for each model in each affordance category, as shown in Fig. 8, where darker
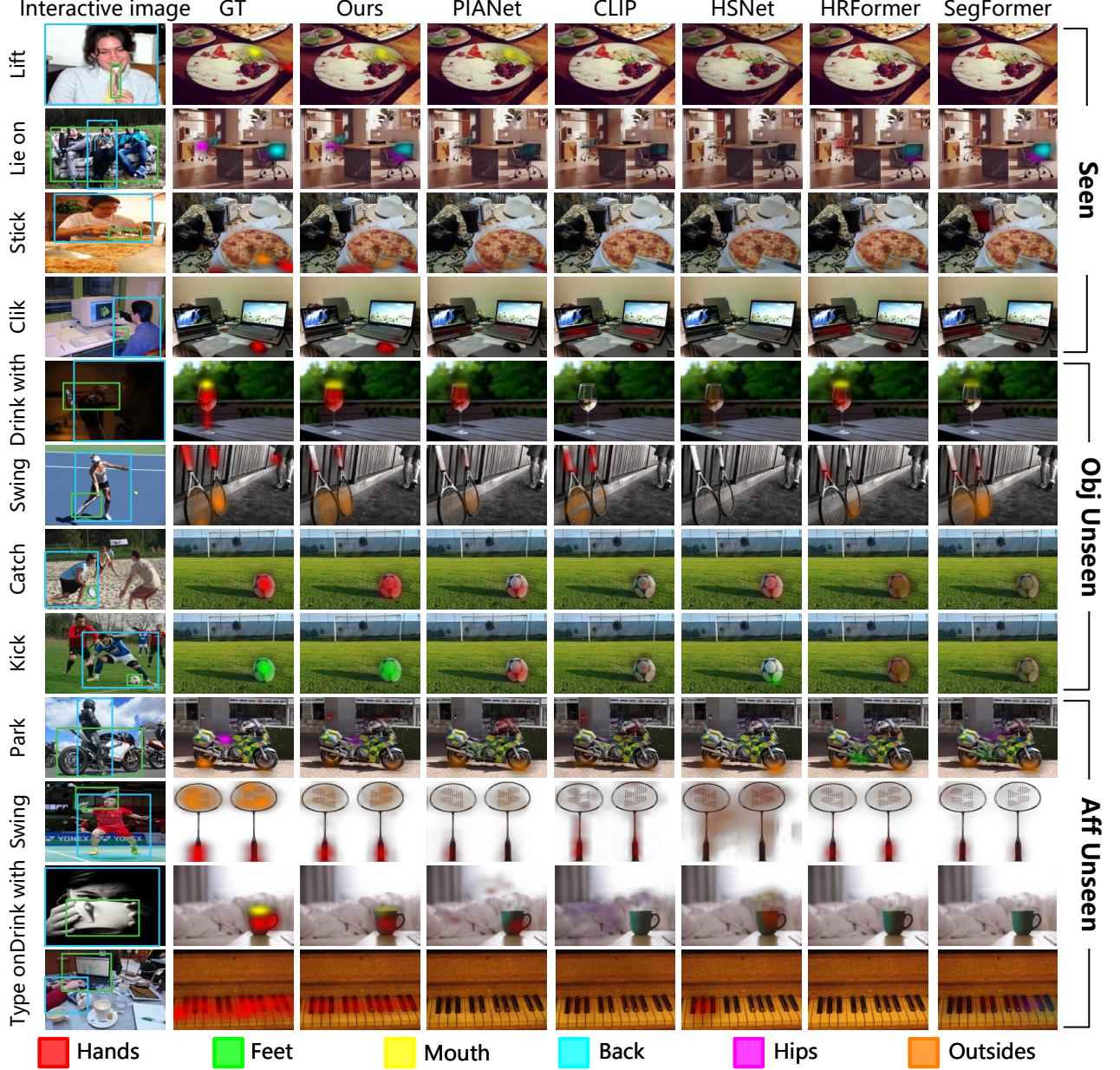
Fig. 9. **Visualization of prediction affordance maps.** We show the visualization results of our model, affordance learning model (PIANet [1]), multimodal model (CLIP [57]), few-shot segmentation model (HSNet [56]), the best human pose estimation model (HRFormer [55]) and the segmentation model (SegFormer [42]).

colours represent a higher ranking of the model's performance under the current category. Our approach is basically the best result in **Seen**, and in **Obj Unseen**, it still achieves better performance for "Park" and "Push", which involve varied and complex scenes with large differences in the appearance of the objects, which suggests that our approach can achieve a more accurate perception of unseen objects based on the appearance and structural similarity of the objects. Our method achieves competitive performance in most of categories in **Aff Unseen** setting, and in the "Drink with" and "Sip" categories involving bottles, cups, and wine glasses, the performance of our method is more impressive than other methods, which indicates that our model can suppress irrelevant backgrounds in the scene by mining the interactive affinity.

Different body parts. We tested the prediction results for different body parts to study the model's capability to perceive the interaction regions corresponding to different body parts, as shown in Table III. Our model achieves almost the best results at the **Seen** setting. Our method still maintains optimal or sub-optimal results on most metrics at the **Obj Unseen** setting. Under the setting of **Aff Unseen**, our model retains very competitive performance on most metrics, among which, for cases like "Mouth" where the contact region is small and occluded, our model has a considerable advantage over other methods, which may be because we consider the use of a large language model to guide the network to explicitly focus on the contact regions of different body parts, eliminating the multiple body parts' ambiguity triggered when they act on the object at the same time, thus achieving more accurate predictions.

TABLE III
DIFFERENT PARTS. WE EVALUATE THE PREDICTIONS FOR EACH BODY PART AND OBJECT CONTACT REGION. THE **BOLD** AND <u>UNDERLINE</u> REPRESENT THE BEST AND SUB-OPTIMAL RESULTS, RESPECTIVELY.

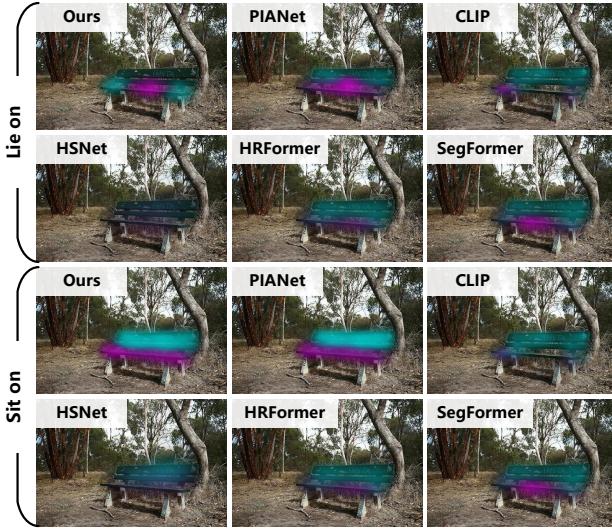|	Method	Hand			Feet			Mouth			Hips			Back			Eyes			Outsides		
Seen	SegFormer	1.263	0.629	3.002	0.923	0.610	3.240	1.442	0.633	3.247	0.929	0.609	2.663	0.703	0.677	3.980	1.136	0.680	2.719	0.898	0.691	3.278
	HFormer	0.859	0.593	2.904	0.768	0.575	3.036	1.028	0.610	3.238	0.748	0.578	2.555	0.599	0.646	3.828	0.785	0.677	2.748	0.673	0.658	3.175
	HSNet	1.511	0.341	1.426	1.367	0.384	1.771	1.800	0.268	1.184	1.277	0.377	1.601	1.460	0.324	1.795	1.500	0.351	1.162	1.254	0.424	1.887
	CLIP	0.935	0.524	2.806	0.896	0.509	2.790	0.971	0.529	3.234	0.821	0.547	2.524	0.855	0.536	3.438	0.690	0.645	2.844	0.818	0.562	2.950
	PIANet	0.669	0.673	3.320	0.607	0.657	3.564	0.865	0.656	3.465	0.600	0.660	2.966	0.548	0.677	3.988	0.684	0.697	2.914	0.541	0.717	3.473
	Ours	0.663	0.677	3.400	0.588	0.656	3.589	0.799	0.673	3.615	0.548	0.689	3.118	0.496	0.694	4.015	0.657	0.705	2.930	0.497	0.722	3.525
Obj Unseen	SegFormer	1.490	0.504	2.361	0.889	0.551	2.992	1.261	0.557	3.072	1.507	0.461	1.844	1.313	0.484	3.183	0.620	0.625	1.957	0.935	0.620	2.394
	HFormer	1.671	0.490	2.237	1.074	0.509	2.754	1.759	0.508	2.812	1.415	0.464	1.929	1.570	0.422	2.607	1.064	0.578	1.659	1.147	0.590	2.306
	HSNet	1.663	0.310	1.215	1.437	0.354	1.873	1.755	0.268	1.031	1.436	0.334	1.351	1.697	0.273	1.834	1.081	0.428	1.136	1.205	0.431	1.564
	CLIP	1.398	0.431	2.097	1.180	0.442	2.458	1.097	0.492	2.859	1.112	0.478	2.125	1.601	0.366	2.602	0.768	0.566	1.912	0.972	0.524	2.133
	PIANet	1.388	0.507	2.423	1.245	0.508	2.730	0.893	0.567	3.151	1.238	0.520	2.223	1.173	0.509	3.430	2.080	0.428	1.186	0.877	0.627	2.558
	Ours	1.262	0.513	2.542	0.885	0.548	3.158	1.131	0.534	3.019	1.046	0.528	2.387	1.121	0.492	3.331	0.980	0.517	1.414	0.774	0.635	2.707
Aff Unseen	SegFormer	2.243	0.310	1.317	0.798	0.568	1.425	2.788	0.262	0.998	2.230	0.343	1.072	1.548	0.441	2.227	2.813	0.308	0.881	1.484	0.458	1.845
	HFormer	2.429	0.329	1.409	1.207	0.519	1.372	3.988	0.215	0.692	2.958	0.292	0.846	2.052	0.399	1.906	3.277	0.294	0.895	1.631	0.462	1.804
	HSNet	1.774	0.279	1.130	0.459	0.673	2.079	1.971	0.229	0.785	1.464	0.334	1.210	1.601	0.293	1.326	1.765	0.288	1.171	1.431	0.371	1.361
	CLIP	1.720	0.303	1.541	0.977	0.505	1.496	2.341	0.217	0.694	1.582	0.350	1.032	1.355	0.380	2.068	2.321	0.340	1.342	1.189	0.435	1.904
	PIANet	1.618	0.357	1.719	0.696	0.578	2.043	2.281	0.245	1.075	1.654	0.320	0.965	1.530	0.346	1.742	2.625	0.286	1.047	1.166	0.468	2.004
	Ours	1.607	0.343	1.726	0.556	0.636	1.768	1.924	0.292	1.435	1.679	0.344	1.028	1.337	0.424	2.291	2.551	0.309	1.103	1.143	0.479	2.024

Fig. 10. **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".
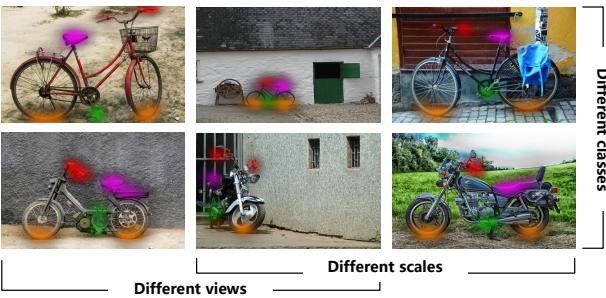

Fig. 11. **Different interactive images** w.r.t. **Same non-interactive images.** We present the prediction results of the model for different non-interactive images of the same interactive image.

Fig. 12. **F-measure curves and PR curves of** 9 **models on the CAL dataset.** The left and right are the F-curve and PR-curve under different settings, respectively.

Affordance's ambiguous uncertainty. Fig. 10 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 interaction variations and accurately transfer the int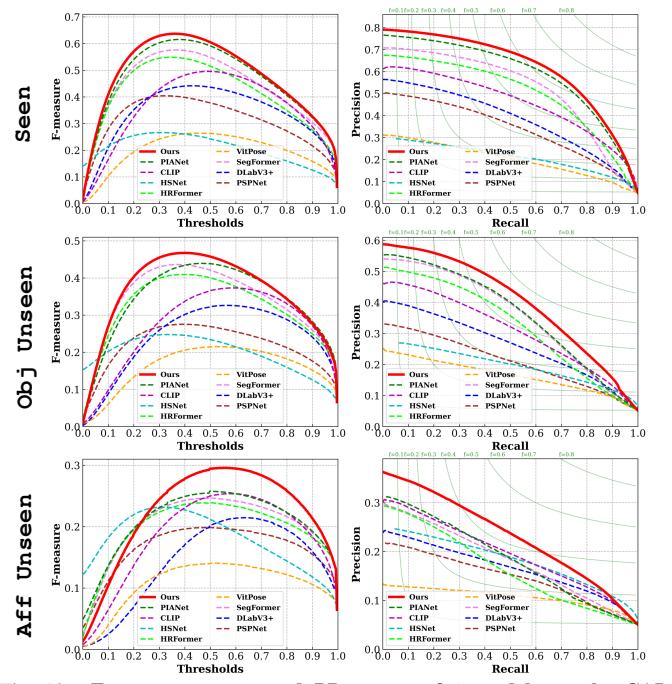eraction-related invariant features to the corresponding regions. Fig. 11 shows the prediction for the same interactive image corresponding to different non-interactive images. Although different object categories, various scales, and different views, our model can transfer the interactive affinity correctly for each part, which indicates that the interactive affinity extraction module can effectively establish the connection between the interactable regions from different branches and facilitate the transfer of interactive affinity.

F-curve and P-R-curve. The F-measure and PR curves are shown in Fig. 12. Our method's PR and F-measure curve are at the top in most situations, indicating that it achieves the best performance on average. And it is even more pronounced in the `Aff Unseen` setting, which suggests that our method has a greater advantage than other approaches when dealing

	Seen			Obj Unseen			Aff Unseen		
	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
w/o text	0.613	0.687	3.518	1.259	0.540	2.629	1.641	0.365	1.599
w/o pose	0.664	0.683	3.458	1.219	0.544	2.625	1.635	0.366	1.615
w/o app	0.708	0.665	3.385	1.252	0.532	2.565	1.624	0.351	1.548
Ours	**0.595**	**0.692**	**3.556**	**1.056**	**0.555**	**2.756**	**1.489**	**0.382**	**1.711**

Fig. 13. **Visualization results for ablation study.** We visualize the results of ablation study for the text, pose and apparent similarity.

with unseen interactions and objects.

D. Ablation Study

In order to verify the effectiveness of our approach to jointly mine interactive affinity from both semantic as well as structural similarity and the necessity of transferring affinity cues based on geometric and apparent similarity, we conducted the corresponding ablation experiments and visualized some of the predictions, as shown in Table IV and Fig. 13. It shows that text has a more negligible effect on model performance, while apparent similarity has a larger influence on the model, suggesting a larger relationship between the possible interactions exhibited by similar local regions. From the visualization results, it is evident that text enables the model to more accurately localize the region of the object that the person is interacting with (*e.g.*, Fig. 13 (a)). The human pose can assist the network to more accurately perceive the relationship between the object structure and the body, and accurately reason about the interactable regions corresponding to different objects (*e.g.*, Fig. 13 (b)). Object apparent similarity enables the model to accurately perceive and locate all affordance regions in complex backgrounds containing multiple object images (*e.g.*, Fig. 13 (c)). Fig. 14 shows the effect of semantics and pose on the interactive interaction region in the SHP module. This shows that semantics explicitly guides the network to focus on different body parts, while pose guides the network to focus on interaction-related object regions, thereby assisting the network to more accurately mine the interactive affinity.

To explore the dynamic adaptation of the DEQ fusion layer to varying inputs, we visualize the statistics of the number of iterations of the DEQ fuse layer and the results of $\|\boldsymbol{Z} -$
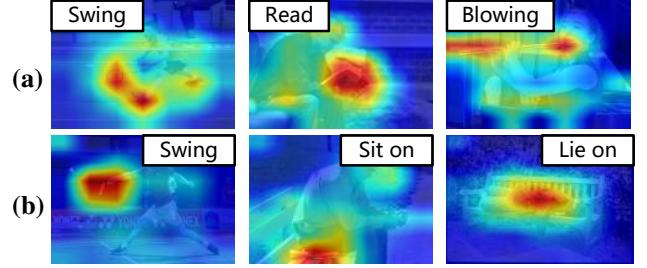

Fig. 14. **Visualization for semantic and pose guidance of interaction regions.** (a) Effect of semantics on interactive feature in the SPH module. (b) Effect of pose on interactive feature in the SPH module
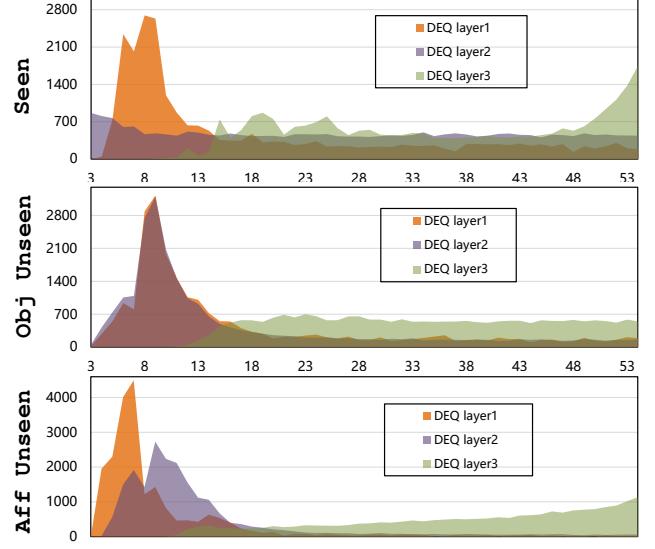

Fig. 15. **Number of DEQ fusion layer iterations.** These represent the results of the statistical distribution of the number of iterations under different settings, respectively.

$f_\theta(\boldsymbol{X}, \boldsymbol{Z})\|$, as shown in Fig. 15 and Fig. 16. It shows that there is a major discrepancy in the statistical distribution of the number of iterations of the DEQ fusion layer in the all settings, which indicates that it is hard to learn the complex dependency between the pose and the image with a fixed number of layers of the network, whereas the DEQ fusion layer can adaptively adjust the computation according to the inputs of the different branches, which makes the model more stable and mine the corresponding local feature representations. In the Fig. 16, the output of Z converges to a stable state quickly, *i.e.*, the DEQ fusion layer finds a stable and efficient balance, ensuring the stability of the model inference. Fig. 17 shows the results of other multi-source feature fusion approaches. For either a fixed f_θ or a fixed transformer layer, it is difficult for the model to adaptively adjust the interactions between different pieces of information, leading to pose which makes it difficult to direct the model's focus on the structural cues related to the interactions, thus hindering the model's performance.

VI. CONCLUSION AND DISCUSSION

This paper proposes to leverage interactive affinity for effective affordance learning by counteracting the influence of multiple possibilities. To this end, a isual-geometric collab-
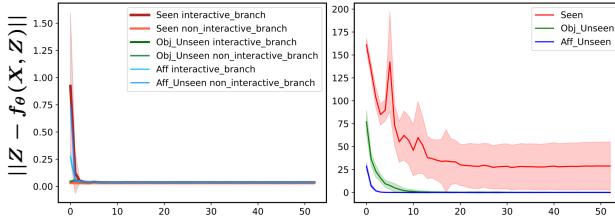
Fig. 16. **Convergence of fixed points.** We show computational convergence results for different DEQ layer indeterminate points.
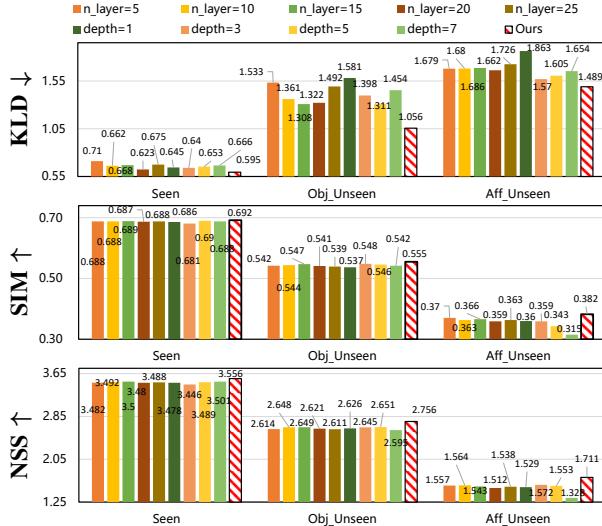

Fig. 17. **Different formats of multi-source feature fusion modules.** We investigate other alternatives to the DEQ layer, where $n\_layer = N_1$ means that the input data goes through the same transformer layer N_1 times, and $depth = N_2$ represents that the input data passes through the N_2 layer transformer for feature fusion.

orative guided affordance learning is introduced, which can 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 $55,047$ images from 35 affordance categories. Our model outperforms eight representative models in three related fields and can serve as a strong baseline for future affordance learning research.

Weakness. Despite our approach achieving excellent prediction results on affordance learning and a strong generalization, there are still some shortcomings with the model's performance relying more on the similarity of the local representations. This paper introduces pose as a bridge to mine geometrically similar cues, which still perform a relatively smaller role than appearance. We will consider considering cues on geometric structures more explicitly in 3D to achieve more accurate perception.

Future Directions. We then discuss several promising future research directions: **1) Affordance learning from egocentric videos.** The first-person video can focus more on interaction-related cues [62], understand the intent of human actions [63, 64], and capture interaction details more accurately, enabling finer-grained affordance learning. **2) Combination of large multimodal models.** The existing large multimodal models can understand the content in the images and reason

accordingly. In the future, we will consider the chain-of-thought method [65] to further reason about the content in interactive images and extract the interactive affinity representation more accurately. **3) Collaborative relationship perception.** Human-object interactions are often associated with the collaboration of both hands/feet. The collaborative action between different body parts can more accurately infer the human-object interaction and the object-object interactions in accomplishing a certain task. Thus, future research should incorporate synergies between different human body parts to understand and fully describe human-object interactions.

REFERENCES

[1] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Leverage interactive affinity for affordance learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6809–6819.

[2] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.

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

[4] T. Nagarajan and K. Grauman, "Learning affordance landscapes for interaction exploration in 3d environments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2005–2015, 2020.

[5] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1164–1172.

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

[7] N. Yamanobe, W. Wan, I. G. Ramirez-Alpizar, D. Petit, T. Tsuji, S. Akizuki, M. Hashimoto, K. Nagata, and K. Harada, "A brief review of affordance in robotic manipulation research," *Advanced Robotics*, vol. 31, no. 19-20, pp. 1086–1101, 2017.

[8] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722–2730.

[9] Y. Sun, S. Fang, and Z. J. Zhang, "Impression management strategies on enterprise social media platforms: An affordance perspective," *International Journal of Information Management*, vol. 60, p. 102359, 2021.

[10] M. Hassanin, S. Khan, and M. Tahtali, "Visual affordance and function understanding: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–35, 2021.

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

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

[13] J. Sawatzky and J. Gall, "Adaptive binarization for weakly supervised affordance segmentation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1383–1391.

[14] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2765–2770.

[15] G. Schiavi, P. Wulkop, G. Rizzi, L. Ott, R. Siegwart, and J. J. Chung, "Learning agent-aware affordances for closed-loop interaction with articulated objects," *arXiv preprint arXiv:2209.05802*, 2022.

[16] C. Ning, R. Wu, H. Lu, K. Mo, and H. Dong, "Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] Y. Wang, R. Wu, K. Mo, J. Ke, Q. Fan, L. J. Guibas, and H. Dong, "Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions," in *European conference on computer vision*. Springer, 2022, pp. 90–107.

[18] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "One-shot affordance detection," *arXiv preprint arXiv:2106.14747*, 2021.

[19] ——, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2252–2261.

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

[21] F. Yang, W. Chen, K. Yang, H. Lin, D. Luo, C. Tang, Z. Li, and Y. Wang, "Learning granularity-aware affordances from human-object interaction for tool-based functional grasping in dexterous robotics," *arXiv preprint arXiv:2407.00614*, 2024.

[22] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," *Advances in neural information processing systems*, vol. 32, 2019.

[23] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao, "One-shot object affordance detection in the wild," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2472–2500, 2022.

[24] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Grounded affordance from exocentric view," *International Journal of Computer Vision*, pp. 1–25, 2023.

[25] X. Wang, R. Girdhar, and A. Gupta, "Binge watching: Scaling affordance learning from sitcoms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2596–2605.

[26] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, "3d affordancenet: A benchmark for visual object affordance understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1778–1787.

[27] J. Sawatzky, A. Srikantha, and J. Gall, "Weakly supervised affordance detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2795–2804.

[28] Y.-C. Liao, K. Todi, A. Acharya, A. Keurulainen, A. Howes, and A. Oulasvirta, "Rediscovering affordance: A reinforcement learning perspective," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–15.

[29] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7081–7091.

[30] H. Geng, S. Wei, C. Deng, B. Shen, H. Wang, and L. Guibas, "Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions," *arXiv preprint arXiv:2312.01307*, 2023.

[31] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 061–18 070.

[32] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 10 922–10 931.

[33] Y. Yang, W. Zhai, H. Luo, Y. Cao, J. Luo, and Z.-J. Zha, "Grounding 3d object affordance from 2d interactions in images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 10 905–10 915.

[34] J. Chen, D. Gao, K. Q. Lin, and M. Z. Shou, "Affordance grounding from demonstration video to target image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 6799–6808.

[35] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 381–389.

[37] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 74–83.

[38] S. Shimada, V. Golyanik, Z. Li, P. Pérez, W. Xu, and C. Theobalt, "Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance," in *European Conference on Computer Vision*. Springer, 2022, pp. 516–533.

[39] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object

[40] W. Mao, M. Liu, R. Hartley, and M. Salzmann, "Contact-aware human motion forecasting," *arXiv preprint arXiv:2210.03954*, 2022.

[41] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu, "Cpf: Learning a contact potential field to model the hand-object interaction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 097–11 106.

[42] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[44] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962v2*, 2019.

[45] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, "Smpler-x: Scaling up expressive human pose and shape estimation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[48] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[49] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," 2015.

[50] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[51] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[53] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[54] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.

[55] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," 2021.

[56] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

[58] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[59] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision (IJCV)*, vol. 7, no. 1, pp. 11–32, 1991.

[60] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[62] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[63] Z. Zhang, H. Luo, W. Zhai, Y. Cao, and Y. Kang, "Pear: Phrase-based hand-object interaction anticipation," *arXiv preprint arXiv:2407.21510*, 2024.

[64] ——, "Bidirectional progressive transformer for interaction intention anticipation," *arXiv preprint arXiv:2405.05552*, 2024.

[65] J. Tang, G. Zheng, J. Yu, and S. Yang, "Cotdet: Affordance knowledge prompting for task driven object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3068–3078.