Grounding 3D Object Affordance from 2D Interactions in Images

Yuhang Yang[1], Wei Zhai[1,*], Hongchen Luo[1], Yang Cao[1,2], Jiebo Luo[3], Zheng-Jun Zha[1]
[1] University of Science and Technology of China [3] University of Rochester
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yyuhang@mail., wzhai056@mail., lhc12@mail., forrest@}ustc.edu.cn
jluo@cs.rochester.edu zhazj@ustc.edu.cn

Abstract

Grounding 3D object affordance seeks to locate objects' "action possibilities" regions in the 3D space, which serves as a link between perception and operation for embodied agents. Existing studies primarily focus on connecting visual affordances with geometry structures, e.g.*, relying on annotations to declare interactive regions of interest on the object and establishing a mapping between the regions and affordances. However, the essence of learning object affordance is to understand how to use it, and the manner that detaches interactions is limited in generalization. Normally, humans possess the ability to perceive object affordances in the physical world through demonstration images or videos. Motivated by this, we introduce a novel task setting: grounding 3D object affordance from 2D interactions in images, which faces the challenge of anticipating affordance through interactions of different sources. To address this problem, we devise a novel Interaction-driven 3D Affordance Grounding Network (IAG), which aligns the region feature of objects from different sources and models the interactive contexts for 3D object affordance grounding. Besides, we collect a Point-Image Affordance Dataset (PIAD) to support the proposed task. Comprehensive experiments on PIAD demonstrate the reliability of the proposed task and the superiority of our method. The project is available at https://github.com/yyvhang/IAGNet.*

1. Introduction

The term "affordance" is described as "opportunities of interaction" by J. Gibson [14]. Grounding 3D object affordance aims to comprehend the interactive regions of objects in 3D space, which is not only simply to predict which interaction an object affords, but also to identify specific points on the object that could support the interaction. It constitutes a link between perception and operation for embodied agents, which has the potential to serve numerous practical applications, *e.g.* action prediction [21, 63], robot manipulation [36, 45, 51], imitation learning [17, 50], and augmented/virtual reality [6, 9].

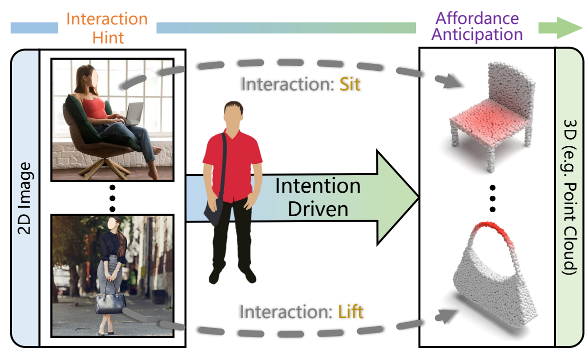

Figure 1. **Grounding Affordance from Interactions.** We propose to ground 3D object affordance through 2D interactions. Given an object point cloud with an interactive image, grounding the corresponding affordance on the 3D object.

So far, the paradigm of perceiving 3D object affordance has several branches. One of them involves establishing an explicit mapping between affordance categories and geometry structures [10, 20, 46, 66], based on visual appearance. However, affordance is dynamic and multiple, these geometric-specific manners have limited generalization for unseen structures. Besides, locking the geometry with a specific affordance category may lead to the anticipated region being inconsistent with its affordance when objects possess multiple similar geometrics, resulting in affordance regional confusion. Another paradigm is based on reinforcement learning, which puts the agent in 3D synthetic scenarios to interact with several objects actively, taking the reward mechanism to optimize the whole process [48]. While this type of approach transitions agents from passive recognition to active reasoning, it needs repeated attempts in a huge search space when meeting a novel structure, and is therefore time-consuming.

*Corresponding Author.

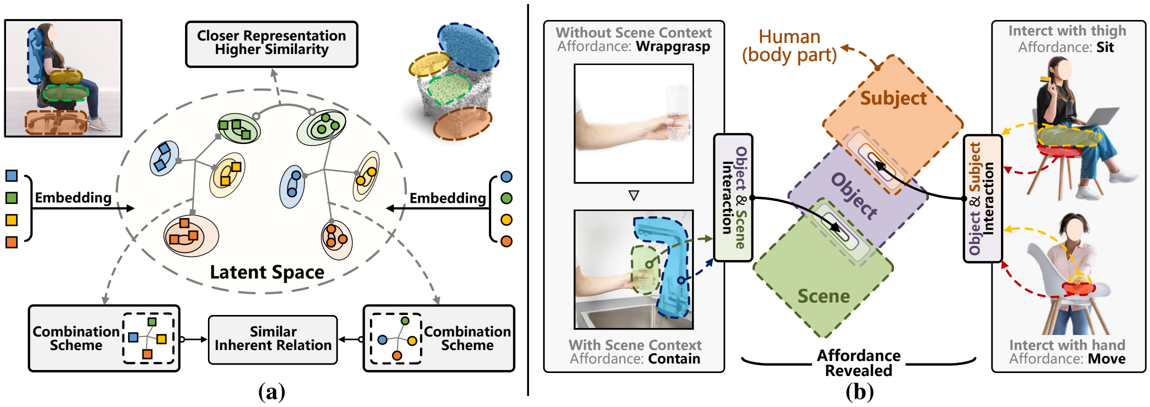

Figure 2. **Motivation.** **(a)** Clues to correlate regions of objects from different sources. Objects with the same category possess a similar combination scheme to meet certain affordance properties, it is a similar inherent relation among different instances. And some structures hint at these properties exhibit closer representations with higher similarity in the latent space. **(b)** Object affordance may be affected by dynamic factors, such as the position of the object itself in the scene, other objects in the scene, and the body part of the interacting subject. These factors can be decomposed into object-subject and object-scene interaction contexts.

These limitations motivate us to explore an affordance-compatible learning paradigm. Typically, humans can infer object affordance in the 3D physical world by watching images or videos that demonstrate the interactions. Some studies in cognitive science [37, 56] point out the existence of "body image" in human cognition, which claims that humans have a perceptual experience of the objects they see, thus facilitating their ability to organize the perception [73] and operate novel objects. Hence, human-object interaction conveys the perception that object structure could perform certain affordance, which is a crucial clue to reason object affordance. In light of this, we present a novel task setting: grounding 3D object affordance from 2D interactions in images, which is shown in Fig. 1.

This challenging task includes several essential issues that should be properly addressed. **1) Alignment ambiguity**. To ground 3D object affordance from 2D source interactions, the premise is to correspond the regions of the object in different sources. The object in 2D demonstrations and the 3D object we face are usually derived from different physical instances in different locations and scales. This discrepancy may lead to confusion when matching the affordance regions, causing alignment ambiguity. While objects are commonly designed to satisfy certain needs of human beings, so the same category generally follows a similar combination scheme of object components to meet certain affordances, and these affordances are hinted at by some structures (Fig. 2 (a)). These invariant properties are across instances and could be utilized to correlate object regions from different sources. **2) Affordance ambiguity**. Affordance has properties of dynamic and multiplicity, which means the object affordance may change according to the situation, the same part of an object could afford multiple interactions, as shown in Fig. 2 (b), "Chair" affords "Sit" or "Move" depends on the human actions, "Mug" affords "Wrapgrasp" or "Contain" according to the scene context, these properties may make ambiguity when extracting affordance. However, these dynamic factors for affordance extraction can be decomposed into the interaction between subject-object and object-scene. Modeling these interactions is possible to extract explicit affordance.

To address these issues, we propose the **I**nteraction-driven 3D **A**ffordance **G**rounding Network (**IAG**) to align object region features from different sources and model interaction contexts to reveal affordance. In detail, it contains two sub-modules, one is **J**oint **R**egion **A**lignment Module (**JRA**), which is devised to eliminate alignment ambiguity. It takes the relative difference in dense cross-similarity to refine analogous shape regions and employs learnable layers to map the invariant combination scheme, taking them to match local regions of objects. For affordance ambiguity, the other one **A**ffordance **R**evealed **M**odule (**ARM**) takes the object representation as a shared factor and jointly models its interaction contexts with the affordance-related factors to reveal explicit affordance. Moreover, we collect **P**oint-**I**mage **A**ffordance **D**ataset (**PIAD**) that contains plenty of paired image-point cloud affordance data and make a benchmark to support the model training and evaluation on the proposed setting.

The contributions are summarized as follows: **1)** We introduce grounding 3D object affordance through the 2D interactions, which facilitates the generalization to 3D object affordance perception. **2)** We propose the IAG framework, which aligns the region feature of objects from different sources and jointly models affordance-related interactions to locate the 3D object affordance. **3)** We collect a dataset named PIAD to support the proposed task setting, which contains paired image-point cloud affordance data. Besides, we establish a benchmark on PIAD, and the experiments on PIAD exhibit the reliability of the method and the setting.

2. Related Works

2.1. Affordance Learning

At present, the affordance area has made achievements in multiple tasks (Tab. 1). Some works are devoted to detecting the affordance region from 2D sources *i.e*. images and videos [7, 11, 25, 33, 54, 60, 78], and there are also some studies accomplish this task with the assistance of natural language [32, 38, 39]. They seek to detect or segment objects that afford the action in 2D data, but cannot perceive the specific parts corresponding to the affordance. Thus, another type of method brings a leap to locate the affordance region of objects from 2D sources [12, 35, 47, 34]. However, the affordance knowledge derived from 2D sources is hard to extrapolate specific interactive locations of objects in the 3D environment. With several 3D object datasets proposed [13, 29, 44], some researchers explore grounding object affordance from 3D data [10, 43, 66, 48]. Such methods directly map the connection between semantic affordances and 3D structures, detach the real interaction, and may deter the generalization ability: structures that do not map to a specific affordance are usually hard to generalize through this type of method. T. Nagarajan *et al*. [48] give a fresh mind, taking the reinforcement learning to make agents actively interact with the 3D synthetic scenarios, while it requires repeated attempts by the agents in a given search space, and is time-consuming. In robotics, affordance is utilized to provide priors for object manipulation and achieve considerable results, especially the articulated 3D objects [42, 64, 79]. Several methods utilize the 2.5D data *e.g*. RGB-D image to understand the object affordance and serve the manipulation task [21, 22, 23, 49], for these methods, the image and depth information are corresponding in spatial and need to be collected in the same scene. In contrast, our task focus on grounding 3D object affordance from 2D interactions, in which the 2D interactions provide direct clues to excavate object affordance efficiently and make the affordance could generalize to some unseen structures, and the 2D-3D data is collected from different sources, freeing the constraints on spatial correspondence.

2.2. Image-Point Cloud Cross-Modal Learning

The combination of point cloud and image data can capture both semantic and geometric information, enabling cross-modal learning among them has great potential application value in scenarios like autopilot [8]. For this type of work, aligning features is the premise for completing downstream tasks *e.g*. detection, segmentation, and registration. It is aimed at establishing correspondences between instances from different modalities [2, 26, 69], either spatial or semantic. To achieve this, many methods use camera intrinsics to correspond spatial position of pixels and points, then align per pixel-point feature or fuse the raw data [24, 59, 62, 68, 77, 80]. Some works utilize depth information to project image features into 3D space and then fuse them with point-wise features [18, 41, 72, 75, 76]. The above methods rely on the spatial prior information to align features, while in our task, images and point clouds are collected from distinct physical instances, there are no priors on camera pose or intrinsics, also no corresponding depth information of the image. Therefore, we align the object region features that are derived from different sources in the feature space with the assistance of the correlation invariance between affordances and appearance structures.

Table 1. **Affordance Learning.** Various works for several tasks in the affordance community.

Work	Input	Output	Task
[11, 60, 78]	Image/Video	2D Mask	Detection
[32, 38, 39]	Image,Text	2D Mask	Detection
[21, 23, 49]	RGBD	Heatmap & Action	Manipulation
[42, 64, 79]	Point Cloud	Heatmap & Action	Manipulation
[12, 35, 47]	Image/Video	2D Heatmap	Grounding
[10, 43, 66]	Point Cloud	3D Heatmap	Grounding

3. Method

3.1. Overview

Our goal is to anticipate the affordance regions on the point cloud that correspond to the interaction in the image. Given a sample $\{P, I, \mathcal{B}, y\}$, where P is a point cloud with the coordinates $P_c \in \mathbb{R}^{N\times 3}$ and the affordance annotation $P_{label} \in \mathbb{R}^{N\times 1}$, $I \in \mathbb{R}^{3\times H\times W}$ is an RGB image, $\mathcal{B} = \{B_{sub}, B_{obj}\}$ denotes the bounding box of the subject and object in I, and y is the affordance category label. The **IAG** (Fig. 3) capture localized features $\mathbf{F_I} \in \mathbb{R}^{C\times H'\times W'}$ and $\mathbf{F}_p \in \mathbb{R}^{C\times N_p}$ of the image and point cloud by two feature extractors ResNet [16] and PointNet++ [52]. Then, utilizing B_{obj}/B_{sub} to locate the object/subject region in $\mathbf{F_I}$, outside the B_{obj} and B_{sub} is the scene mask M_{sce}, use ROI-Align [15] to get the object, subject, and scene features $\mathbf{F}_i, \mathbf{F}_s, \mathbf{F}_e \in \mathbb{R}^{C\times H_1\times W_1}$, and reshape them to $\mathbb{R}^{C\times N_i}$ ($N_i = H_1 \times W_1$). Next, the JRA module takes $\mathbf{F}_i, \mathbf{F}_p$ as input and jointly mines the correspondence to align them, obtaining the joint feature $\mathbf{F}_j$ (Sec. 3.2). Following, the ARM module takes $\mathbf{F}_j, \mathbf{F}_s, \mathbf{F}_e$ to reveal the affordance $\mathbf{F}_\alpha$ through cross-attention mechanism (Sec. 3.3). Ultimately, $\mathbf{F}_\alpha$ and $\mathbf{F}_j$ are sent to the decoder to compute the affordance logits $\hat{y}$ and the final 3D object affordance $\hat{\phi}$ (Sec. 3.4). The process is expressed as $\hat{\phi}, \hat{y} = f_\theta(P_c, I, \mathcal{B}; \theta)$, where f_θ denotes the network and θ is the parameter.

3.2. Joint Region Alignment Module

The JRA module calculates the high-dimensional dense similarity in feature space to approximate analogous shape

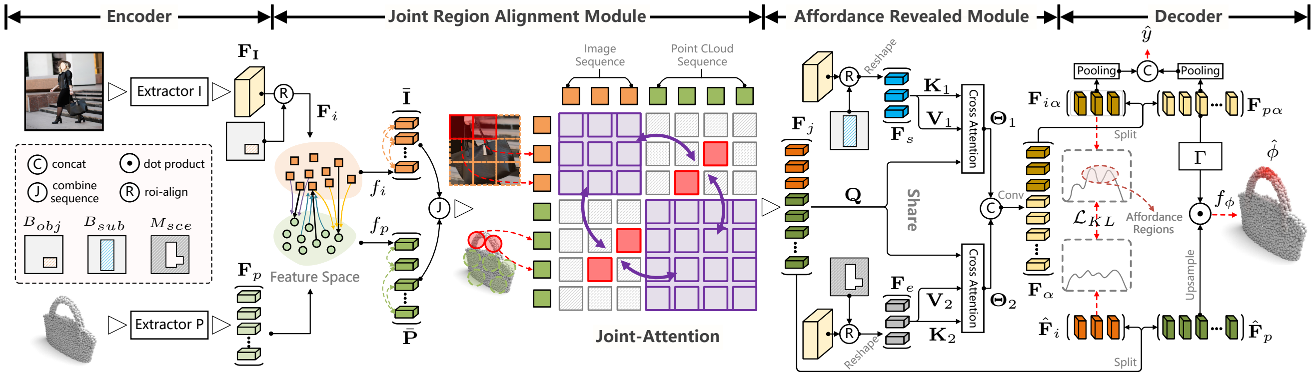

Figure 3. **Method.** Overview of Interaction-driven 3D Affordance Grounding Network (IAG), it firstly extracts localized features $\mathbf{F}_i, \mathbf{F}_p$ respectively, then takes JRA (Sec. 3.2) to align them and get the joint feature $\mathbf{F}_j$. Next, ARM (Sec. 3.3) utilizes $\mathbf{F}_j$ to reveal affordance $\mathbf{F}_\alpha$ with $\mathbf{F}_s, \mathbf{F}_e$ by cross-attention. Eventually, $\mathbf{F}_j$ and $\mathbf{F}_\alpha$ are sent to the decoder (Sec. 3.4) to obtain the final results $\hat{\phi}$ and $\hat{y}$.

regions that contain relatively higher similarity [2]. Meanwhile, to map the inherent combination scheme of object components, the JRA models the inherent relations between modality-specific regions and takes it to drive the alignment of other regions by a transformer-based joint-attention f_δ. With the network trained to capture the correspondences among regions from different sources, the alignment is performed implicitly during the optimization.

Initially, $\mathbf{F}_p$ and $\mathbf{F}_i$ are projected into a feature space by shared convolution layers f_σ, obtaining the features $\mathbf{P} \in \mathbb{R}^{C\times N_p}, \mathbf{I} \in \mathbb{R}^{C\times N_i}$. Then, to correlate the analogous shape regions, the dense cross-similarity is calculated between each region of $\mathbf{P}$ and $\mathbf{I}$, formulated as:

$$\varphi_{i,j} = \frac{e^{(\mathbf{P}_i,\mathbf{I}_j)}}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_i} e^{(\mathbf{P}_i,\mathbf{I}_j)}}, \varphi \in \mathbb{R}^{N_p\times N_i}, \tag{1}$$

where $\varphi_{i,j}$ denotes the cross-similarity between the i-th region of $\mathbf{P}$ and the j-th region of $\mathbf{I}$. Taking the relative difference in φ to refine and correlate analogous regions in $\mathbf{P}$ and $\mathbf{I}$, next, applying self-attention layers f_i, f_p to model the intra-structural inherent relation of objects in respective modalities, the process is expressed as:

$$\bar{\mathbf{P}} = f_p(\mathbf{I}\cdot\varphi^T), \bar{\mathbf{I}} = f_i(\mathbf{P}\cdot\varphi), \tag{2}$$

where $\bar{\mathbf{P}} \in \mathbb{R}^{C\times N_p}, \bar{\mathbf{I}} \in \mathbb{R}^{C\times N_i}$. Following, we perform a joint-attention on features with structural relevance to map the similar inherent combination scheme and drive the alignment of the remaining regions. The process is formulated as: $\mathbf{F}_j = f_\delta[\bar{\mathbf{P}}, \bar{\mathbf{I}}]$, where $\mathbf{F}_j \in \mathbb{R}^{C\times(N_p+N_i)}$ and $[\cdot]$ denotes joining the image feature sequence and point cloud feature sequence into a whole one, and $\mathbf{F}_j$ denotes the joint representation. The $\mathbf{F}_j$ is sent to the affordance revealed module to excavate interaction contexts with $\mathbf{F}_s, \mathbf{F}_e$. And the affordance knowledge revealed by the interaction contexts could mutually optimize the alignment process during training (see Sec. 3.4).

3.3. Affordance Revealed Module

In general, the dynamic factors related to affordance are reflected in the interaction, affecting the extraction of affordance representation, and the interaction is mostly considered to exist between humans and objects. However, the interaction between objects and scenes (including other objects) may also affect affordance [43]. To consider both of these factors, the ARM module utilizes the cross-attention technique to extract these interaction contexts respectively and jointly embed them to reveal explicit object affordance.

Specifically, $\mathbf{F}_j$ is projected to form the shared query $\mathbf{Q} = \mathbf{F}_j\mathbf{W}_1$, $\mathbf{F}_s$ and $\mathbf{F}_e$ are projected to form different keys and values $\mathbf{K}_{1/2} = \mathbf{F}_s\mathbf{W}_{2/3}, \mathbf{V}_{1/2} = \mathbf{F}_e\mathbf{W}_{4/5}$, where $\mathbf{W}_{1\sim5}$ are projection weights. Then cross-attention aggregates them to excavate interaction contexts, expressed as:

$$\mathbf{\Theta}_{1/2} = softmax(\mathbf{Q}^T\cdot\mathbf{K}_{1/2}/\sqrt{d})\cdot\mathbf{V}_{1/2}^T, \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{d\times(N_p+N_i)}, \mathbf{K}_{1/2}, \mathbf{V}_{1/2} \in \mathbb{R}^{d\times N_i}$, d is the dimension of projection, $\mathbf{\Theta}_1, \mathbf{\Theta}_2$ indicate the excavated interaction contexts. Then, jointly embedding $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ to reveal affordance, expressed as:

$$\mathbf{F}_\alpha = f_\xi(\mathbf{\Theta}_1, \mathbf{\Theta}_2), \mathbf{F}_\alpha \in \mathbb{R}^{C\times(N_p+N_i)}, \tag{4}$$

where $\mathbf{F}_\alpha$ is the joint affordance representation, it is split in the decoder to compute final results, f_ξ denotes the concatenation, followed by a convolution layer to fuse the feature.

3.4. Decoder and Loss Function

The $\mathbf{F}_j$ is split to $\hat{\mathbf{F}}_p \in \mathbb{R}^{C\times N_p}$ and $\hat{\mathbf{F}}_i \in \mathbb{R}^{C\times N_i}$, $\mathbf{F}_\alpha$ is split to $\mathbf{F}_{p\alpha} \in \mathbb{R}^{C\times N_p}$ and $\mathbf{F}_{i\alpha} \in \mathbb{R}^{C\times N_i}$ in the decoder. Pooling $\mathbf{F}_{p\alpha}$ and $\mathbf{F}_{i\alpha}$ respectively and concatnate them to compute the affordance logits $\hat{y}$. $\hat{\mathbf{F}}_p$ is upsampled to $\mathbb{R}^{C\times N}$ by Feature Propagation Layers (FP) [52], and the 3D object affordance $\hat{\phi}$ is computed as:

$$\hat{\phi} = f_\phi(\mathbf{FP}(\hat{\mathbf{F}}_p)\odot\Gamma(\mathbf{F}_{p\alpha})), \hat{\phi} \in \mathbb{R}^{N\times1}, \tag{5}$$

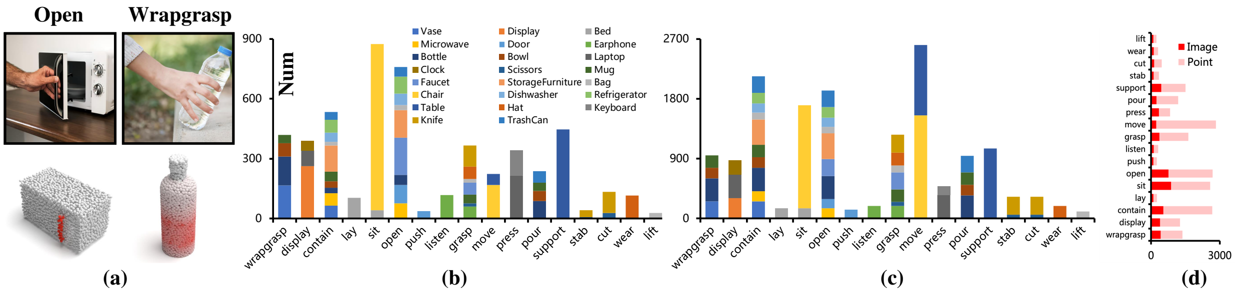

Figure 4. **Properties of the PIAD dataset.** **(a)** Data pairs in the PIAD, the red region in point clouds is the affordance annotation. **(b)** Distribution of the image data. The horizontal axis represents the category of affordance, the vertical axis represents quantity, and different colors represent different objects. **(c)** Distribution of the point cloud data. **(d)** The ratio of images and point clouds in each affordance class. It shows that images and point clouds are not fixed one-to-one pairing, they can form multiple pairs. Please refer to supp. for more details.

where f_ϕ is an output head, $\mathbf{FP}$ is the upsample layer, Γ denotes pooling $\mathbf{F}_{p\alpha}$ and expand it to the shape of $\mathbb{R}^{C\times N}$.

The total loss comprises three parts: $\mathcal{L}_{CE}$, $\mathcal{L}_{KL}$, $\mathcal{L}_{HM}$. Where $\mathcal{L}_{CE}$ computes the cross-entropy loss between y and $\hat{y}$, it supervises the extraction of $\mathbf{F}_\alpha$ and implicitly optimizes the alignment process. To enable the network focus on the alignment of affordance regions, we apply the KL Divergence (KLD) [3] to constrain the distribution between $\mathbf{F}_{i\alpha}$ and $\hat{\mathbf{F}}_i$, formulated as $\mathcal{L}_{KL} = (\mathbf{F}_{i\alpha}||\hat{\mathbf{F}}_i)$. The reason is that $\mathbf{F}_{i\alpha}$ exhibits the affordance distribution of each object region in the image, and the affordance-related regions keep more significant features. Constraining the feature distribution of $\hat{\mathbf{F}}_i$ to enhance the affordance region features in $\hat{\mathbf{F}}_i$, and with the region correlations established by the alignment process, $\hat{\mathbf{F}}_p$ also tends to exhibit this property, similar to distillation [19, 71]. Which makes the alignment and affordance extraction optimize mutually. $\mathcal{L}_{HM}$ is a focal loss [28] combined with a dice loss [40], it is calculated by $\hat{\phi}$ and P_{label}, which supervise the point-wise heatmap on point clouds. Eventually, the total loss is formulated as:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{CE} + \lambda_2\mathcal{L}_{KL} + \lambda_3\mathcal{L}_{HM}, \tag{6}$$

where λ_1, λ_2 and λ_3 are hyper-parameters to balance the total loss. See more details in supplementary materials.

4. Dataset

Collection Details. We collect **P**oint-**I**mage **A**ffordance **D**ataset (**PIAD**), which contains paired image-point cloud affordance data. The point clouds are mainly collected from 3D-AffordanceNet [10], including the point cloud coordinates and affordance annotation. Images are mainly collected from HICO [4], AGD20K [35], and websites with free licenses. The collection criteria is that the image should demonstrate interactions that the object in the point cloud could afford. For example, if the object point cloud is a "Chair", it affords "Sit", then the image should depict a subject (usually humans) sitting on a chair. The final dataset comprises 7012 point clouds and 5162 images, spanning 23 object classes and 17 affordance categories. Notably, objects in the images and point clouds do not sample from the same physical instance, but they belong to the same object category. Paired examples are shown in Fig. 4 (a).

Annotation Details. For point clouds, each point may afford one or multiple interactions, *e.g.* one annotation is a matrix of (2048, 17), 2048 is the point number, 17 represents the number of affordance types. Each element in the matrix indicates the probability of a point affording a specific affordance. Meanwhile, images are annotated with bounding boxes of the interactive subject and object, as well as an affordance category label. In our task, we only use the heatmap in the point cloud annotation that corresponds to the affordance of the image for training, resulting in a matrix of (2048, 1). By doing so, the affordance category is detached from the point cloud during inference, and the anticipation of the 3D object affordance category only relies on 2D interactions. As a result, different affordances can be anticipated on a point cloud through distinct interactions.

Statistic Analysis. Since images and point clouds are sampled from different instances, they do not need a fixed one-to-one pairing, one image could be paired with multiple point clouds and the count of them is not strictly consistent. Fig. 4 (b) and (c) show the count and distribution of affordances in images and point clouds. Fig. 4 (d) illustrates the ratio of images and point clouds in each affordance category. PIAD has two partitions: **`Seen`** and **`Unseen`**. In **`Seen`**, both objects and affordances in the training and testing sets are consistent, while in **`Unseen`**, some objects or affordances in the testing set do not exist in the training set.

5. Experiments

5.1. Benchmark Setting

Evaluation Metrics. To provide a comprehensive and effective evaluation, we compare serval advanced works in the affordance area [10, 47, 74] and finally chose four evaluation metrics: **AUC** [31], **aIOU** [53], **SIM**ilarity [57] and **M**ean **A**bsolute **E**rror [65] to benchmark the PIAD.

Modular Baselines. Since there are no prior works using paired image-point cloud data to ground 3D object

Table 2. **Comparison Results on PIAD.** The overall results of all comparative methods, the best results are in **bold**. **Seen** and **Unseen** are two partitions of the dataset. "**Green**" and "**Orange**" indicate two types of the comparative modular baselines. "**Base.**" represents the baseline. ◇ denotes the relative improvement of our method over other methods. AUC and aIOU are shown in percentage.

	Metrics	Base.	MBDF [59]	PMF [80]	FRCNN [68]	ILN [5]	PFusion [67]	XMF [1]	Ours
Seen	AUC ↑	68.49	74.88◇13.3%	75.05◇13.0%	76.05◇11.5%	75.84◇11.9%	77.50◇9.5%	78.24◇8.4%	**84.85 ± 0.3**
	aIOU ↑	6.85	9.34◇119.5%	10.13◇102.4%	11.97◇71.3%	11.52◇78.0%	12.31◇66.6%	12.94◇58.5%	**20.51 ± 0.7**
	SIM ↑	0.367	0.415◇31.3%	0.425◇28.2%	0.429◇27.0%	0.427◇27.6%	0.432◇26.1%	0.441◇23.6%	**0.545 ± 0.02**
	MAE ↓	0.152	0.143◇31.4%	0.141◇30.5%	0.136◇27.9%	0.137◇28.4%	0.135◇27.4%	0.127◇22.8%	**0.098 ± 0.01**
Unseen	AUC ↑	57.34	58.23◇23.3%	60.25◇19.2%	61.92◇16.0%	59.69◇20.3%	61.87◇16.1%	62.58◇14.8%	**71.84 ± 1.8**
	aIOU ↑	3.95	4.22◇88.4%	4.67◇70.2%	5.12◇55.2%	4.71◇68.8%	5.33◇49.1%	5.68◇39.9%	**7.95 ± 0.8**
	SIM ↑	0.318	0.325◇8.3%	0.330◇6.6%	0.332◇6.0%	0.325◇8.3%	0.330◇6.6%	0.342◇2.9%	**0.352 ± 0.03**
	MAE ↓	0.235	0.213◇40.3%	0.211◇39.8%	0.195◇34.8%	0.207◇38.6%	0.193◇34.2%	0.188◇32.4%	**0.127 ± 0.01**

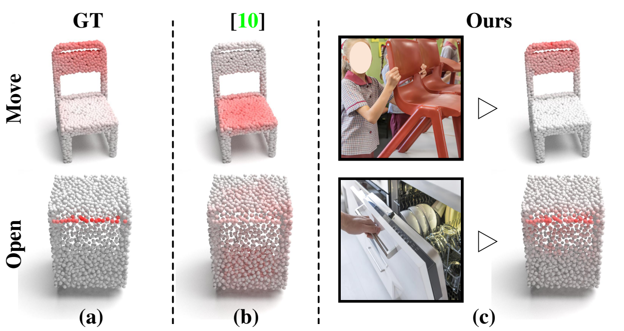

Figure 5. **Paradigm Comparison.** **(a)** Ground truth. **(b)** The results of 3D-AffordanceNet [10]. **(c)** Our results. The top row shows the regional confusion case in **Seen**, and the bottom row displays the generalization in **Unseen**.

Table 3. **Ablation Study.** We investigate the improvement of JRA and ARM on the model performance based on the baseline.

	JRA	ARM	AUC	aIOU	SIM	MAE
Seen			69.92	8.85	0.427	0.132
	✓		80.29	14.31	0.495	0.121
		✓	78.67	13.95	0.475	0.126
	✓	✓	**85.16**	**21.20**	**0.564**	**0.088**
Unseen			59.14	4.05	0.338	0.202
	✓		66.25	6.27	0.363	0.159
		✓	65.79	5.99	0.358	0.162
	✓	✓	**73.69**	**8.70**	**0.383**	**0.117**

affordance. For a thorough comparison of our method, we select several advanced image-point cloud cross-modal works as modular baselines. Methods of comparison are broadly divided into two types, one is that utilizes camera intrinsics to facilitate cross-modal feature alignment or fusion, *i.e.* **MBDF-Net (MBDF)** [59], **PMF** [80] and **FusionRCNN (FRCNN)** [68], for this type of method, we remove the step of using intrinsic parameters to align raw data or features to explore their effectiveness on PIAD. Another type performs feature alignment directly in the feature space, without relying on camera intrinsic, *i.e.* **ImLoveNet (ILN)** [5], **PointFusion (PFusion)** [67] and **XMFnet (XMF)** [1]. To ensure a fair comparison, all methods use the same feature extractors, with the only variation coming from cross-modal alignment or fusion block. For the **Baseline**, we directly concatenate the features that are output from extractors, regarding the concatenation as the cross-fusion block. More details about baselines are provided in the supplementary materials.

5.2. Comparison Results

The comparison results of evaluation metrics are shown in Tab. 2. As can be seen, our method outperforms the compared baselines, across all metrics in both partitions. Besides, to display the limitation of methods that lock geometrics with specific semantic affordance categories, we conduct an experiment to compare one of these methods [10] with ours. As shown in Fig. 5, the top row indicates that the result of this method exhibits regional confusion, *e.g.* the region where the chair could be moved or sat is geometrically rectangular, and it directly anticipates results on these similar geometries, inconsistent with affordance property. Plus, the bottom row shows that directly establishing a link between object structures and affordances may fail to anticipate correct 3D affordance in **Unseen**. In contrast, our method anticipates precise results by mining affordance clues provided by 2D interactions. Additionally, visual comparative results of our method and other baselines are shown in Fig. 6. As can be seen, the comparative baselines could anticipate some 3D object affordance under our setting, but in comparison, our method obviously achieves better results, which validates the rationality of our setting, and also demonstrates the superiority of our method.

5.3. Ablation Study

Effectiveness of JRA. Tab. 3 reports the impact on evaluation metrics of JRA. Fig. 7 provides visual comparison results. It shows that without the JRA, the result is anticipated over the entire region that contains the interactive compo-

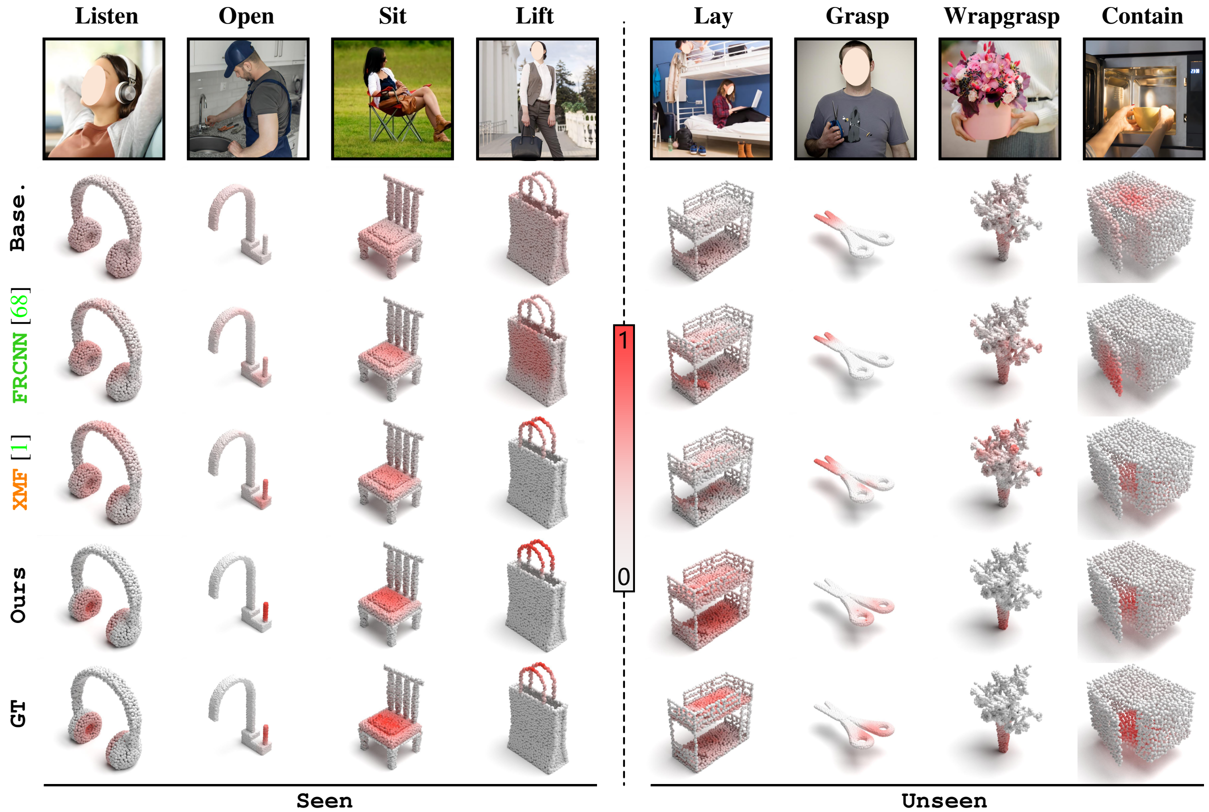

Figure 6. **Visualization Results.** The first row is the interactive image, which demonstrates the interaction that the object can afford. The last row is the ground truth of 3D object affordance in the point cloud. The four columns on the left are the visual comparison results for distinct 3D object affordances in the `Seen` partition. The four columns on the right are the results in the `Unseen` partition.

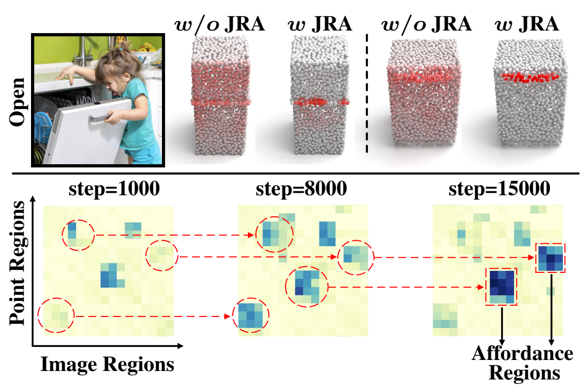

Figure 7. **Ablation of JRA.** The top row is the result of the anticipated affordance with/without JRA. The bottom row is a part of the cross-similarity matrix of a sample during the training process.

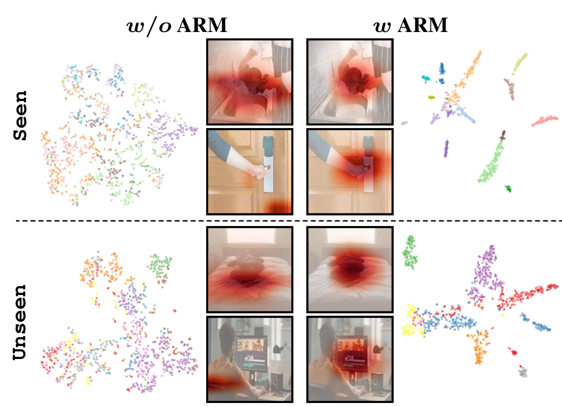

Figure 8. **Ablation of ARM.** The activation map and t-SNE [61] results with (w), without (w/o) ARM in both partitions.

nents, which means the 2D-3D affordance regions of objects do not match well. Besides, we visualize a part of the cross-similarity matrix between $\hat{\mathbf{F}}_p$ and $\hat{\mathbf{F}}_i$ in Fig. 7. In the initial stage, only a few analogous shape regions keep explicit correspondence, as training proceeds, the JRA maps the correlation among regions with lower similarity. Meanwhile, the model extracts explicit affordance progressively, and the affordance introduced into the optimization process reveals the corresponding affordance regions.

Effectiveness of ARM. The influence of ARM on evaluation metrics is also shown in Tab. 3. To visually evaluate the effectiveness of ARM, we employ GradCAM [55] to generate the activation map, and utilize t-SNE [61] to demonstrate the clustering of affordances in both partitions. The results are shown in Fig. 8. It shows that ARM makes the model focus on the interactive regions to excavate the interaction contexts, and enables the model to differentiate various affordances from interactions in both partitions.

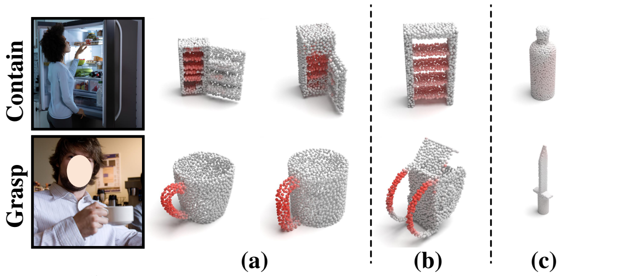

Figure 9. **Same Image** $w.r.t.$ **Multiple Point Clouds.** **(a)** Same object category. **(b)** Different object categories, similar geometrics. **(c)** Different object categories and geometrics.

Figure 10. **Same Point Cloud** $w.r.t.$ **Multiple Images.** Grounding affordance on the same point cloud with images that contain similar or disparate interactions.

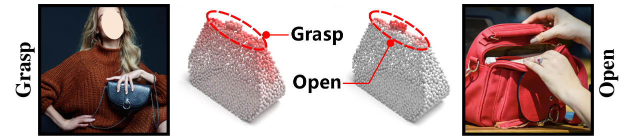

Figure 11. **Multiple Affordances.** Some objects like "Bag" contains the region that corresponds to multiple affordances.

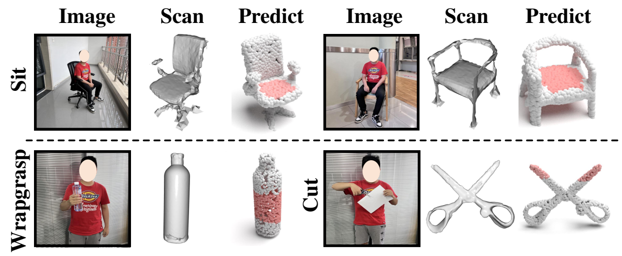

Figure 12. **Real-Wrold.** The first row is scanned by iPhone, the second row comes from [29], point clouds are sampled from them.

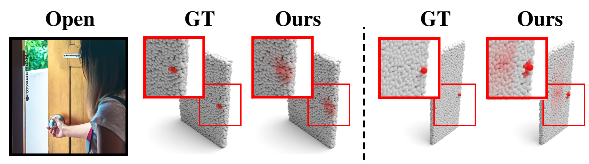

Figure 13. **Failure Cases.** Over prediction in the small regions.

5.4. Performance Analysis

Different Instances. We conduct experiments to verify whether the model could ground 3D affordance on instances from different sources: (i) using an image and different point clouds to infer respectively (Fig. 9 (a)). (ii) using multiple images and a single point cloud (Fig. 10). The results indicate that the model can anticipate 3D affordances on different instances with the same 2D interaction, and can also anticipate distinct 3D affordances on the same point cloud through different 2D interactions. Showing that the model maintains the mapping between object structures and affordances by learning from different 2D interactions.

Different Object Categories. What will happen if the object category is different in the image and point cloud? To explore this issue, we perform experiments with mismatched object categories, shown in Fig. 9 (b) and (c). When objects are of different categories but with similar geometric primitives, *e.g*. the handle of "Mug" and the shoulder strap of "Bag", our model still properly infer the 3D affordance. However, when geometric structures are also dissimilar (Fig. 9 (c)), the model does not make random predictions. This indicates that the model maps cross-category invariance between affordance and geometries, which can be generalized to new instances of geometry.

Multiplicity. One defining property of affordance is multiplicity: some points may correspond to multiple affordances. Fig. 11 shows that the model makes objects' multiple affordances compatible, and verifies the model does not anticipate affordance by linking a region with specific affordances. Instead, it predicts by considering the presence of affordance-related interactions in the same region.

Real World. To validate the model in real-world scenarios, we use an iPhone to scan objects and real-scan dataset [29] to test our model, shown in Fig. 12. It shows that our model exhibits a certain degree of generalization to the real world.

Limitations. Our model exhibits over predictions for some small affordance regions, shown in Fig. 13. This may be due to the limited understanding of fine-grained geometric parts, and aligning the object features in a single scale may lead to overly large receptive fields. Our feature extension will refer [13, 27] to tackle this problem.

6. Conclusion

We present a novel setting for affordance grounding, which utilizes the 2D interactive semantics to guide the grounding of 3D object affordance, it has the potential to serve embodied systems when collaborating with multi-modal grounding systems [30, 58, 70]. Besides, We collect the PIAD dataset as the first test bed for the proposed setting, it contains paired image-point cloud affordance data. Plus, we propose a novel framework to correlate affordance regions of objects that are from different sources and model interactive contexts to ground 3D object affordance. Comprehensive experiments on PIAD display the reliability of the setting, and we believe it could offer fresh insights and facilitate research in the affordance area.

Acknowledgments This work is supported by National Key R&D Program of China under Grant 2020AAA0105700, National Natural Science Foundation of China (NSFC) under Grants 62225207, U19B2038 and 62121002.

References

[1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. *arXiv preprint arXiv:2209.09552*, 2022. 6, 7

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 3, 4

[3] 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

[4] 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*, pages 1017–1025, 2015. 5

[5] Honghua Chen, Zeyong Wei, Yabin Xu, Mingqiang Wei, and Jun Wang. Imlovenet: Misaligned image-supported registration network for low-overlap point cloud pairs. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 6

[6] Kun-Hung Cheng and Chin-Chung Tsai. Affordances of augmented reality in science learning: Suggestions for future research. *Journal of science education and technology*, 22(4):449–462, 2013. 1

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

[8] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):722–739, 2021. 3

[9] Barney Dalgarno and Mark JW Lee. What are the learning affordances of 3-d virtual environments? *British Journal of Educational Technology*, 41(1):10–32, 2010. 1

[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. 1, 3, 5, 6

[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. 3

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

[13] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He 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 (CVPR)*, pages 7081–7091, June 2023. 3, 8

[14] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[17] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017. 1

[18] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[19] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Amalgamating knowledge from heterogeneous graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15709–15718, 2021. 5

[20] David Inkyu Kim and Gaurav S Sukhatme. Semantic labeling of 3d point clouds with object affordance for robot manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5578–5584. IEEE, 2014. 1

[21] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 1, 3

[22] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *European Conference on Computer Vision*, pages 831–847. Springer, 2014. 3

[23] 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. 3

[24] Georg Krispel, Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1874–1883, 2020. 3

[25] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023. 3

[26] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023. 3

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 8

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[29] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14809–14818, June 2022. 3, 8

[30] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019. 8

[31] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008. 5

[32] Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, and Yang Cao. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence*, 2022. 3

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

[34] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *arXiv preprint arXiv:2208.13196*, 2022. 3

[35] 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. 3, 5

[36] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176. IEEE, 2021. 1

[37] Maurice Merleau-Ponty. Phénoménologie de la perception (1945). *Libraqire Gallimard, Paris*, 1976. 2

[38] Jinpeng Mi, Hongzhuo Liang, Nikolaos Katsakis, Song Tang, Qingdu Li, Changshui Zhang, and Jianwei Zhang. Intention-related natural language grounding via object affordance detection and intention semantic extraction. *Frontiers in Neurorobotics*, 14:26, 2020. 3

[39] Jinpeng Mi, Song Tang, Zhen Deng, Michael Goerner, and Jianwei Zhang. Object affordance based multimodal fusion for natural human-robot interaction. *Cognitive Systems Research*, 54:128–137, 2019. 3

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[41] Niluthpol Chowdhury Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Rgb2lidar: Towards solving large-scale cross-modal visual localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 934–954, 2020. 3

[42] 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

[43] 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, 4

[44] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[45] Bogdan Moldovan, Plinio Moreno, Martijn Van Otterlo, José Santos-Victor, and Luc De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *2012 ieee international conference on robotics and automation*, pages 4373–4378. IEEE, 2012. 1

[46] 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

[47] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 3, 5

[48] 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

[49] 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. 3

[50] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018. 1

[51] François Osiurak and Arnaud Badets. Tool use and affordance: Manipulation-based versus reasoning-based approaches. *Psychological review*, 123(5):534, 2016. 1

[52] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3, 4

[53] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016. 5

[54] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *European conference on computer vision*, pages 186–201. Springer, 2016. 3

[55] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[56] Chris Shilling. *The body in culture, technology and society*. Sage, 2004. 2

[57] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991. 5

[58] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*, 2020. 8

[59] Xun Tan, Xingyu Chen, Guowei Zhang, Jishiyu Ding, and Xuguang Lan. Mbdf-net: Multi-branch deep fusion network for 3d object detection. In *Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data*, pages 9–17, 2021. 3, 6

[60] Spyridon Thermos, Petros Daras, and Gerasimos Potamianos. A deep learning approach to object affordance segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2358–2362. IEEE, 2020. 3

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[62] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 3

[63] Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, and Josef Sivic. Predicting actions from static scenes. In *European Conference on Computer Vision*, pages 421–436. Springer, 2014. 1

[64] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 90–107. Springer, 2022. 3

[65] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005. 5

[66] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 1, 3

[67] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 244–253, 2018. 6

[68] Xinli Xu, Shaocong Dong, Lihe Ding, Jie Wang, Tingfa Xu, and Jianan Li. Fusionrcnn: Lidar-camera fusion for two-stage 3d object detection. *arXiv preprint arXiv:2209.10733*, 2022. 3, 6, 7

[69] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement, 2023. 3

[70] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2561–2569, 2019. 8

[71] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 5

[72] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021. 3

[73] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. Exploring figure-ground assignment mechanism in perceptual organization. *Advances in Neural Information Processing Systems*, 35:17030–17042, 2022. 2

[74] 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. 5

[75] Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, and Kai Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4534–4543, 2020. 3

[76] Linqing Zhao, Jiwen Lu, and Jie Zhou. Similarity-aware fusion network for 3d semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1585–1592. IEEE, 2021. 3

[77] Lin Zhao, Hui Zhou, Xinge Zhu, Xiao Song, Hongsheng Li, and Wenbing Tao. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *arXiv preprint arXiv:2108.07511*, 2021. 3

[78] Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18):14321–14333, 2020. 3

[79] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *arXiv preprint arXiv:2207.01971*, 2022. 3

[80] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021. 3, 6