Learning 2D Invariant Affordance Knowledge for 3D Affordance Grounding

Xianqiang Gao[*1,2], **Pingrui Zhang**[*2,3], **Delin Qu**[2,3]
Dong Wang[2], **Zhigang Wang**[2], **Yan Ding**[2], **Bin Zhao**[2,4†]
[1]University of Science and Technology of China
[2]Shanghai AI Laboratory
[3]Fudan University
[4]Northwestern Polytechnical University
gaoxianqiang@mail.ustc.edu.cn, {zhangpingrui,zhaobin}@pjlab.org.cn

Abstract

3D Object Affordance Grounding aims to predict the functional regions on a 3D object and lays the foundation for a wide range of applications in robotics. Recent advances tackle this problem via learning a mapping between 3D regions and a single human-object interaction image. However, the geometric structure of the 3D object and the object in the human-object interaction image are not always consistent, leading to poor generalization. To address this issue, we propose to learn generalizable invariant affordance knowledge from multiple human-object interaction images within the same affordance category. Specifically, we introduce the **M**ulti-**I**mage Guided Invariant-**F**eature-Aware 3D **A**ffordance **G**rounding (**MIFAG**) framework. It grounds 3D object affordance regions by identifying common interaction patterns across multiple human-object interaction images. First, the Invariant Affordance Knowledge Extraction Module (**IAM**) utilizes an iterative updating strategy to gradually extract aligned affordance knowledge from multiple images and integrate it into an affordance dictionary. Then, the Affordance Dictionary Adaptive Fusion Module (**ADM**) learns comprehensive point cloud representations that consider all affordance candidates in multiple images. In addition, the Multi-Image and Point Affordance (**MIPA**) benchmark is constructed and our method outperforms existing state-of-the-art methods in various experimental comparisons.

Code — https://github.com/goxq/MIFAG-code

1 Introduction

3D Object Affordance Grounding seeks to identify and predict functional regions on an object's 3D point cloud. This task has laid the foundation for connecting visual perception with physical operation and paved the way for a wide range of real-world applications, such as embodied systems (Ahn et al. 2022; Driess et al. 2023; Wu et al. 2023), object manipulation (Huang et al. 2024; Li et al. 2024b; Huang et al. 2023; Wu et al. 2021) and object grasping (Dai et al. 2023).

Currently, methods in 3D object affordance prediction can be divided into two categories. One of them involves utilizing reinforcement learning, which requires an agent to interact with objects repeatedly in a simulated environment. Such methods are usually time-consuming, and their generalization performance is often restricted by the limited number of simulation scenes (Mo et al. 2021; Ning et al. 2024; Cheng et al. 2023). The other category leverages labeled 3D object affordance data and attempts to learn a mapping between objects and affordance in an end-to-end manner. This kind of approach is more direct and has gained more and more popularity recently. Specifically, 3D AffordanceNet (Deng et al. 2021) first constructs a 3D point cloud affordance grounding benchmark. After that, to accommodate the diversity of object affordances and incorporate human interaction references, many methods have been proposed, either combining a single image or textual description with the point cloud for 3D affordance prediction (Yang et al. 2023; Li et al. 2024c).

Though pioneer research studies have achieved promising progress, they are still limited by failing to leverage the strong correlations and implied invariant affordance knowledge among objects within the same affordance category. As

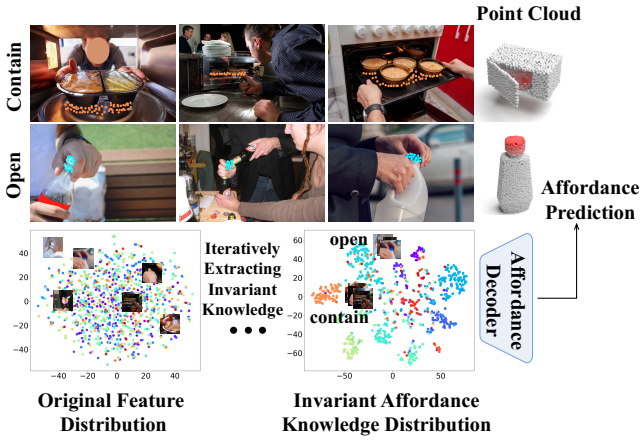


Figure 1: **Motivation of Our Method.** The reference human-object images in each row exhibit significant variations in appearance, yet they consistently represent the same affordance knowledge. We propose to iteratively extract the invariant affordance knowledge from multiple images, leading to improved performance.

[*]These authors contributed equally.
[†]Corresponding author

shown in Figure 1, the affordance of an object is usually determined by multiple human-object reference images. These real-world images exhibit significant variations in appearance yet represent the same object and affordance category. Thus, they share invariant affordance knowledge and strong internal relationships, providing valuable insights into the object's affordance across different contexts. Similarly, predicting the affordance of an object requires the invariant knowledge derived from numerous human-object interaction images. However, previous approaches have not fully addressed these limitations, often either lacking visual information or simply focusing on mapping 3D regions to a single human-object interaction image, leading to poor generalization and suboptimal performance.

Overcoming the aforementioned limitations is non-trivial due to the following challenges: (1) In real-world scenarios, objects of the same category often exhibit significant variations in their interactive regions. Relying solely on textual descriptions, such as "open the oven", is insufficient for conveying affordance knowledge. For instance, ovens from different brands can vary in size and handle position. Additionally, simply combining multiple images as references may not yield the expected results due to the considerable diversity in their appearances (*e.g.*, shape, scale, *etc.*). Therefore, the first challenge lies in how to effectively utilize multiple images with diverse appearances for affordance guidance and how to extract the invariant affordance knowledge. (2) There is a significant gap between the modalities of reference images and point clouds, presenting another challenge: how to effectively integrate invariant affordance knowledge into the point cloud for accurate affordance prediction.

To address the above challenges, we propose the **M**ulti-**I**mage Guided Invariant-**F**eature-Aware 3D **A**ffordance **G**rounding (**MIFAG**) framework, which gradually extracts affordance knowledge from multiple human-object reference images and effectively integrates this invariant knowledge with point cloud representations to achieve accurate affordance prediction. Specifically, our proposed MIFAG consists of two modules: the Invariant Affordance Knowledge Extraction Module (**IAM**) and the Affordance Dictionary Adaptive Fusion Module (**ADM**). The IAM progressively extracts invariant affordance knowledge across different images using a multi-layer network, while its dual-branch structure minimizes interference caused by appearance variations in the images. The output of the last layer of IAM forms an affordance dictionary that encapsulates all invariant affordance knowledge from these images. Following this, we design the ADM to fuse the extracted invariant affordance knowledge with point clouds, and use the point cloud to query each candidate in the affordance dictionary, thereby obtaining comprehensive point cloud feature representations that consider all affordance candidates.

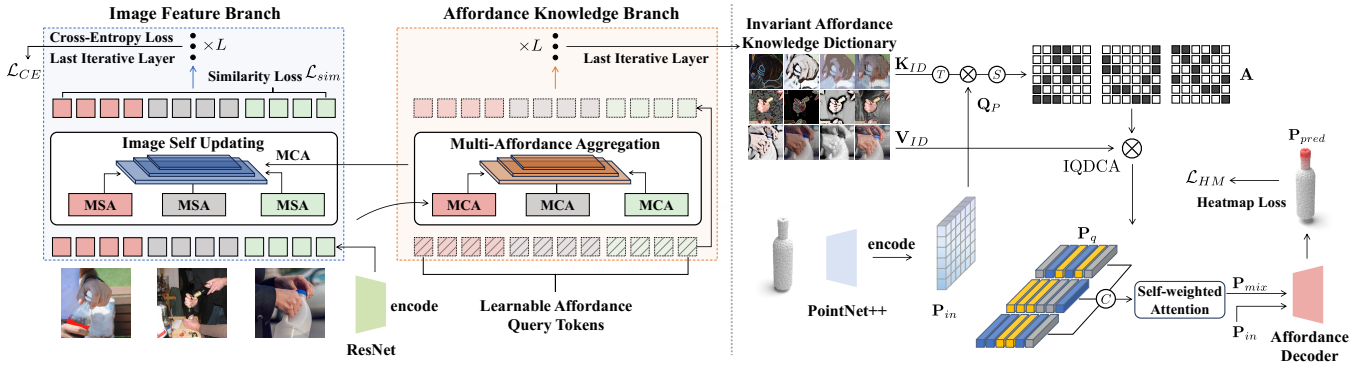
Our main contributions can be summarized as follows:

- We introduce a novel MIFAG framework for 3D object affordance grounding, which extracts invariant affordance knowledge from multiple reference images.
- We propose the IAM, which progressively extracts invariant affordance knowledge from images while minimizing interference caused by appearance variations. The ADM is then proposed to leverage this knowledge to obtain comprehensive point cloud features that consider all affordance candidates.
- We construct the Multi-Image and Point Affordance (MIPA) benchmark to advance research in understanding affordances across diverse visual data. Experimental results demonstrate that our MIFAG outperforms previous state-of-the-art methods.

2 Related Work

2.1 Affordance Learning

Affordance learning is crucial in robotics, particularly for manipulating articulated objects (Ning et al. 2024). Consequently, many studies focus on detecting the affordance areas of target objects or scenes (Luo et al. 2022; Fang et al. 2018). Some works extract affordance knowledge from 2D data, *i.e.*, images and videos (Luo et al. 2021; Chen et al. 2023; Liu et al. 2024), while others explore affordance using natural language guidance (Chen, Cong, and Kan 2024; Guo et al. 2024; Li et al. 2024a; Qian et al. 2024). These methods aim to predict the affordance regions of 2D targets. However, robotics tasks often require 3D information, and transferring affordance knowledge learned from 2D data to 3D scenarios can lead to failures in real-world applications. As a result, several 3D affordance-related datasets have been proposed (Geng et al. 2023b; Mo et al. 2019), and some studies focus on leveraging these 3D datasets to ground object affordance (Xu et al. 2022a; Delitzas et al. 2024). 3D AffordanceNet (Deng et al. 2021) first introduces a benchmark dataset for grounding affordance regions on object point clouds. Building on this 3D dataset, IAGNet (Yang et al. 2023) proposes a method for learning 3D affordance from reference images. However, the significant appearance variations among reference images and 3D point clouds pose a challenge. LASO (Li et al. 2024c) deals with this by replacing images with natural language descriptions of the interaction area. However, completely discarding the vision information may overlook distinct affordance area guidance. Therefore, we propose a method that better aligns visual latent features from multiple images to effectively guide affordance grounding. Where2Explore (Ning et al. 2024) acquires affordance information by explicitly estimating the geometric similarity across different categories, thereby enhancing generalization. GAPartNet (Geng et al. 2023b) introduces cross-category tasks and a dataset to explore the consistency of generalizable and actionable parts. PartManip (Geng et al. 2023a) identifies actionable parts to facilitate cross-category object manipulation. While these approaches do not fully capture the category-level consistency features crucial for affordance grounding, our method improves by explicitly focusing on consistent features across images and more effectively learning affordance information from reference images using dual-branch methods (Yu et al. 2024a,b).



(a) Invariant Affordance Knowledge Extraction Module (IAM) **(b) Affordance Dictionary Adaptive Fusion Module (ADM)**

Figure 2: **Overview of our proposed MIFAG.** (a) The **IAM** utilizes a multi-layer network and a dual-branch structure to gradually extract invariant affordance knowledge and minimize interference caused by appearance variations in the images. (b) The **ADM** leverages the invariant affordance knowledge dictionary derived from (a), using dictionary-based cross attention and self-weighted attention to comprehensively fuse the affordance knowledge with point cloud representations.

2.2 Image-Point Cloud Cross-Modal Learning

Cross-modal tasks enhance individual modalities by incorporating information from one or more other modalities. These additional modalities can serve as conditions to guide the learning progress, often resulting in a positive impact (Qu et al. 2024). The combination of images and point clouds is particularly complementary in downstream tasks: point clouds provide stereospatial information, while images offer rich color and texture details. To leverage this complementarity, many works focus on aligning each pixel to a corresponding point to enrich the semantic information of raw data (Zhuang et al. 2021; Xu et al. 2022b; Tan et al. 2021; Aiello, Valsesia, and Magli 2022; Chen et al. 2022; Xu, Anguelov, and Jain 2018). Additionally, some methods aim to learn multi-view fusion with point clouds, seeking common knowledge and aligning both image-to-image and image-to-point cloud (Jaritz, Gu, and Su 2019; Zhao, Lu, and Zhou 2021; Chen et al. 2024). However, unlike these approaches, which may not fully capture the relationships between multiple images and often lack sufficient modeling of consistent image features for effective fusion with point clouds, our method first establishes consistency between images and then aligns the invariant image features with the point cloud, leading to a more robust integration of visual and spatial information.

3 Method

3.1 Overview

Given a 3D object $\mathbf{P} = \{\mathbf{P}_c, \mathbf{P}_{label}\}$ and its corresponding n reference images $\{I_1, I_2, \ldots, I_n\}$, where $\mathbf{P}_c \in \mathbb{R}^{N\times 3}$ represents the point cloud coordinates, $\mathbf{P}_{label} \in \mathbb{R}^{N\times 1}$ denotes the affordance annotation for each coordinate, and $I_i \in \mathbb{R}^{3\times H\times W}, i = 1, \ldots, n$, our goal is to ground the affordance region $\mathbf{P}_{pred} \in \mathbb{R}^{N\times 1}$ on the point cloud using the invariant affordance knowledge derived from the n images.

As illustrated in Figure 2, our **MIFAG** framework consists of two modules: the Invariant Affordance Knowledge Extraction Module (**IAM**) and the Affordance Dictionary Adaptive Fusion Module (**ADM**), which are used to extract invariant affordance knowledge across multiple images and fuse this knowledge with point clouds, respectively. Specifically, in the IAM, a multi-layer network is employed to gradually extract affordance knowledge from images, while a dual-branch structure is designed to minimize interference caused by appearance variations in the images. The output of the last layer of the Affordance Knowledge Branch forms an Invariant Affordance Knowledge Dictionary that encapsulates the affordance knowledge across all reference images. Then, in the ADM, we apply dictionary-based cross-attention and self-weighted attention to fuse the invariant affordance knowledge with the point clouds, resulting in the output $\mathbf{P}_{mix} \in \mathbb{R}^{n\times N\times C}$. Finally, an affordance decoder is used to produce the final affordance prediction $\mathbf{P}_{pred} \in \mathbb{R}^{N\times 1}$.

3.2 Invariant Affordance Knowledge Extraction Module

The appearances of the object among the reference images vary significantly, yet they all share a common affordance category and provide valuable insights into how the object can be used. Accordingly, we design the IAM to align multiple images based on their common affordance type and to gradually extract the invariant affordance knowledge using a multi-layer network and dual-branch structure.

As illustrated in Figure 2 (a), we first randomly initialize a series of Learnable Affordance Query Tokens $\mathbf{Q}$, which are used to represent the learned invariant affordance knowledge from the reference images. These query tokens are then iteratively updated by the multi-layer network using multi-head cross attention, with the reference image features serving as the key and value, as formulated below:

$$\mathbf{Q}_i^{(l)} = \mathrm{MCA}(\mathbf{Q}_i^{(l-1)}\mathbf{W}_q, \mathbf{F}_i^{(l-1)}\mathbf{W}_k, \mathbf{F}_i^{(l-1)}\mathbf{W}_v), \quad (1)$$

where $\mathbf{Q}_i^{(l-1)} \in \mathbb{R}^{M\times C}$ and $\mathbf{F}_i^{(l-1)} \in \mathbb{R}^{D\times H\times W}$ denote the affordance queries and reference image features from layer

$(l-1)$, respectively. Then, to leverage the inherent consistency among multiple affordance queries, we apply an MLP layer to align and aggregate all the queries. The aggregated features are then fed back into the image feature branch by multi-head cross attention. Additionally, to continually extract consistent image features, the image feature branch is also updated iteratively using a multi-head self-attention layer. This process can be expressed as follows:

$$\mathbf{Q}_f^{(l)} = \mathrm{MLP}(\mathbf{Q}_1^{(l)}, \mathbf{Q}_2^{(l)}, \ldots, \mathbf{Q}_n^{(l)}), \tag{2}$$

$$\bar{\mathbf{F}}_i^{(l-1)} = \mathrm{MSA}(\mathbf{F}_i^{(l-1)}\mathbf{W}), \tag{3}$$

$$\mathbf{F}_i^{(l)} = \mathrm{MCA}(\bar{\mathbf{F}}_i^{(l-1)}\mathbf{W}_q, \mathbf{Q}_f^{(l)T}\mathbf{W}_k, \mathbf{Q}_f^{(l)}\mathbf{W}_v), \tag{4}$$

where $\mathbf{Q}_f^{(l)}, \mathbf{Q}_1^{(l)}, \ldots, \mathbf{Q}_n^{(l)} \in \mathbb{R}^{M\times C}$ and $\bar{\mathbf{F}}_i^{(l-1)}, \mathbf{F}_i^{(l)} \in \mathbb{R}^{D\times H\times W}$. At every iterative layer, to enforce the constraint that all images share the same affordance category, we calculate the similarity loss among all image features $\mathbf{F}_1^{(l)}, \mathbf{F}_2^{(l)}, \ldots, \mathbf{F}_n^{(l)}$.

In the IAM, the affordance knowledge branch and image feature branch work in tandem. The learnable affordance query tokens iteratively interact with the image features, adding useful affordance information to the image feature branch. Meanwhile, in the affordance knowledge branch, invariant affordance knowledge is gradually aggregated under the guidance of the image feature branch. This dual-branch structure minimizes interference from varying image features on the affordance knowledge. Finally, the output of the last iterative layer of the affordance knowledge branch forms the Invariant Affordance Knowledge Dictionary that encapsulates the affordance knowledge from reference images.

3.3 Affordance Dictionary Adaptive Fusion Module

To leverage the Invariant Affordance Knowledge Dictionary obtained in Section 3.2, we design the ADM to adaptively integrate the affordance knowledge into the point cloud representations. As illustrated in Figure 2 (b), we first calculate the point-feature weighted query $\mathbf{P}_q$. Next, $\mathbf{P}_q$ is fused through a self-weighted attention layer to obtain the point-affordance-mixed feature $\mathbf{P}_{mix}$.

Unlike existing multi-head self-attention, which generates query, key, and value tokens from the input feature itself, our approach utilizes the obtained Invariant Affordance Knowledge as an extra dictionary to seamlessly guide affordance learning on the point cloud during the training phase (Zhang et al. 2024). In line with this, we propose the adaptive Invariant-aware Query Dictionary Cross-Attention (IQDCA). Initially, we use the point cloud feature $\mathbf{P}_{in}$ to generate the query $\mathbf{Q}_P$ for cross-attention, while the keys $\mathbf{K}_{q_1}, \mathbf{K}_{q_2}, \ldots, \mathbf{K}_{q_n}$ and values $\mathbf{V}_{q_1}, \mathbf{V}_{q_2}, \ldots, \mathbf{V}_{q_n}$ are obtained from the Invariant Affordance Knowledge Dictionary as follows:

$$\mathbf{Q}_P = \mathbf{P}_{in}\mathbf{W}_q, \tag{5}$$

$$\mathbf{K}_{ID} = [\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n]\mathbf{W}_k, \tag{6}$$

$$\mathbf{V}_{ID} = [\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n]\mathbf{W}_v, \tag{7}$$

where $\mathbf{Q}_P \in \mathbb{R}^{N\times d}, \mathbf{K}_{ID}, \mathbf{V}_{ID} \in \mathbb{R}^{n\times M\times d}$. We then apply the cross-attention mechanism to calculate the dictionary attention matrix $\mathbf{A}$:

$$\mathbf{A} = \mathrm{SoftMax}(\mathrm{Sim}_{\cos}(\mathbf{Q}_P, \mathbf{K}_{ID})), \tag{8}$$

where $\mathrm{Sim}_{\cos}$ denotes the cosine similarity function. The dictionary attention matrix $\mathbf{A} \in \mathbb{R}^{n\times N\times M}$ represents the similarity between the point cloud and the invariant query for each individual image. This matrix serves as guidance to adaptively refine the knowledge in the dictionary. The entire IQDCA process can be expressed as follows:

$$\mathrm{IQDCA}(\mathbf{Q}_P, \mathbf{K}_{ID}, \mathbf{V}_{ID}) = \mathbf{A}\cdot\mathbf{V}_{ID}. \tag{9}$$

The output of the IQDCA is $\mathbf{P}_q \in \mathbb{R}^{n\times N\times d}$, which contains all the weighted invariant queries. Then, we apply a self-weighted attention layer to discard irrelevant tokens to obtain the final affordance dictionary-based adaptive fusion feature $\mathbf{P}_{mix}$:

$$\mathbf{P}_{mix} = \mathrm{SWA}([\mathbf{P}_{q_1}, \mathbf{P}_{q_2}, \ldots, \mathbf{P}_{q_n}]). \tag{10}$$

Next, this adaptively weighted fusion feature is combined with the original point cloud feature $\mathbf{P}_{in}$ to obtain the final feature $\mathbf{P}_{out}$, which is then fed into the affordance decoder to predict the distribution of the affordance region.

3.4 Decoder and Loss Function

We pool the last layer outputs of the IAM, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n$, to compute the affordance logits $\hat{y}$. Additionally, we send $\mathbf{P}_{out}$ to the decoder f_d to ground the 3D affordance $\mathbf{P}_{pred}$:

$$\mathbf{P}_{pred} = f_d(\mathbf{P}_{out}), \tag{11}$$

where $\mathbf{P}_{pred} \in \mathbb{R}^{N\times 1}$. The total loss consists of three components: $\mathcal{L}_{CE}$, $\mathcal{L}_{Sim}$ and $\mathcal{L}_{HM}$. $\mathcal{L}_{CE}$ represents the cross-entropy loss between y and $\hat{y}$. $\mathcal{L}_{Sim}$ is the cosine similarity loss calculated at each layer of $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n\}$, which aims to align image features while obtaining invariant affordance knowledge. $\mathcal{L}_{HM}$ combines focal loss (Lin et al. 2017) with dice loss (Milletari, Navab, and Ahmadi 2016), and is calculated between $\mathbf{P}_{pred}$ and $\mathbf{P}_{label}$, supervising the point-wise heatmap on point clouds. The total loss is formulated as follows:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{CE} + \lambda_2\mathcal{L}_{Sim} + \lambda_3\mathcal{L}_{HM}, \tag{12}$$

where λ_1, λ_2, and λ_3 are hyperparameters used to balance the total loss. Further details can be found in the supplementary materials (Gao et al. 2024).

4 Experiments

4.1 Experimental Settings

Dataset To the best of our knowledge, there is currently no affordance dataset that satisfactorily tackles challenges posed by significant variations in appearance of human-object interaction images. Common datasets lack visual information or primarily focus on a single object corresponding image, ignoring the invariant affordance knowledge and internal relationships across different object contexts. To address these challenges, we constructed the **M**ulti-**I**mage and

	Metrics	PMF	MBDF	FRCNN	ILN	PFusion	XMF	IAGNet	LASO	MIFAG
Seen	AUC ↑	80.46	79.05	80.33	80.48	80.55	80.04	82.95	83.13	**85.10**
	aIOU ↑	10.04	12.68	10.33	10.18	10.78	9.76	17.92	17.27	**20.50**
	SIM ↑	0.445	0.476	0.449	0.447	0.449	0.442	0.547	0.540	**0.568**
	MAE ↓	0.125	0.114	0.124	0.125	0.129	0.129	0.094	0.097	**0.091**
Unseen	AUC ↑	69.14	65.21	68.75	69.66	68.10	68.71	69.91	63.89	**71.13**
	aIOU ↑	3.99	4.65	3.18	4.79	4.46	3.96	5.12	5.18	**5.23**
	SIM ↑	0.302	0.305	0.299	0.304	0.302	0.301	0.310	0.299	**0.315**
	MAE ↓	0.152	0.142	0.213	0.164	0.142	0.172	0.144	0.140	**0.136**

Table 1: **Affordance Prediction Metrics on MIPA Benchmark.** Comparison of evaluation between the proposed method **MIFAG** and baseline methods on **MIPA**. **MIFAG** significantly surpasses existing methods and achieves SOTA performance.

Point **A**ffordance (**MIPA**) Dataset, which comprised paired multi-images and point clouds. We leverage point clouds and affordance annotations from 3D AffordanceNet (Deng et al. 2021), while gathering paired multiple images from IAGNet (Yang et al. 2023), HICO (Chao et al. 2015) and AGD20K (Luo et al. 2022). The proposed **MIPA** dataset contains 5,162 images and 7,012 point clouds, covering 23 object classes and 17 affordance categories. In addition, we conducted our training and evaluation under the **`seen`** and **`unseen`** settings following previous works (Yang et al. 2023; Li et al. 2024c). The **`seen`** setting shares identical object categories in training and evaluation, whereas the **`unseen`** setting utilizes different splits of categories.

Compared Baselines and Evaluation Metrics The most relevant works are IAGNet (Yang et al. 2023) and LASO (Li et al. 2024c), both of which leverage 3D AffordanceNet (Deng et al. 2021) to derive the corresponding 3D shapes. In addition, consistent with IAGNet, we adopt SOTA image-3D cross-modal methods as baselines for a more comprehensive evaluation, encompassing PMF (Zhuang et al. 2021), XMF (Aiello, Valsesia, and Magli 2022), and ILN (Chen et al. 2022), which concentrate on fusing image and point features, as well as MBDF (Tan et al. 2021), FRCNN (Xu et al. 2022b), and PFusion (Xu, Anguelov, and Jain 2018) which are predominantly employed for multi-modal object grounding. We reproduce these methods with the same feature extractors and settings in the original papers to ensure a fair comparison. Besides, we enhance these methods by replacing the single image or text inputs with multiple ones, allowing for a more direct comparison with our dataset. All the evaluation metrics follow previous works: **A**rea **U**nder the **C**urve (**AUC**) (Lobo, Jiménez-Valverde, and Real 2008), **a**verage **I**ntersection **O**ver **U**nion (**aIOU**) (Rahman and Wang 2016), **SIM**ilarity (**SIM**) (Swain and Ballard 1991) and **M**ean **A**bsolute **E**rror (**MAE**) (Willmott and Matsuura 2005).

Implementation Details Following IAGNet (Yang et al. 2023), we employ PointNet++ (Qi et al. 2017) and ResNet-18 (He et al. 2016) as the default 3D and 2D backbone, respectively. We train MIFAG model on a single NVIDIA A100 GPU with a batch size of 64, using the Adam optimizer with a learning rate of 4e-5. More detailed experiment settings can be found in the supplementary materials.

4.2 Quantitative Analysis

Table 1 reports the metrics of the proposed method **MIFAG** compared with other methods on the **MIPA benchmark**. The results demonstrate that the proposed **MIFAG** achieves state-of-the-art performance, significantly surpassing existing methods in 3D point cloud affordance prediction. Specifically, **MIFAG** achieve scores of 85.10 in **AUC** and 20.50 in **aIOU**, outperforming the second-best method, LASO (Li et al. 2024c), and IAGNet (Yang et al. 2023), with an improvement of +1.97 in **AUC** and +2.58 in **aIOU**. Notably, our method also demonstrates impressive affordance capabilities in the unseen setting, outperforming the second-best method, IAGNet by 1.22 in **AUC** and LASO by 0.05 in **aIOU**. Additionally, 3D segmentation and tracking methods such as PMF (Zhuang et al. 2021), XMF (Aiello, Valsesia, and Magli 2022), and MBDF (Tan et al. 2021) perform significantly worse compared to the affordance grounding methods.

4.3 Qualitative Results

Visualization of point cloud affordance grounding. Figure 3 illustrates the point cloud affordance grounding results of various methods in both **`seen`** and **`unseen`** settings of the **MIPA** dataset. The results demonstrate that our method achieves more accurate results, outperforming LASO and IAGNet. For instance, on objects such as the "door" and the "vase", LASO and IAGNet suffer from affordance map dispersion or omission, whereas our approach can accurately focus on the interactive regions.

Qualitative Analysis by t-SNE. In Figure 4, we visualize the distribution of query tokens as they propagate through the IAM network, demonstrating its effectiveness in extracting invariant affordance knowledge. The affordance query tokens are randomly initialized at the beginning. As the affordance query tokens are iteratively updated, they gradually cluster based on affordance type. For instance, features corresponding to the same "open" operation across different objects, such as "Door", "Microwave", and "Bag", cluster in the orange region. This result demonstrates the effectiveness of our approach by precisely extracting invariant knowledge across images.

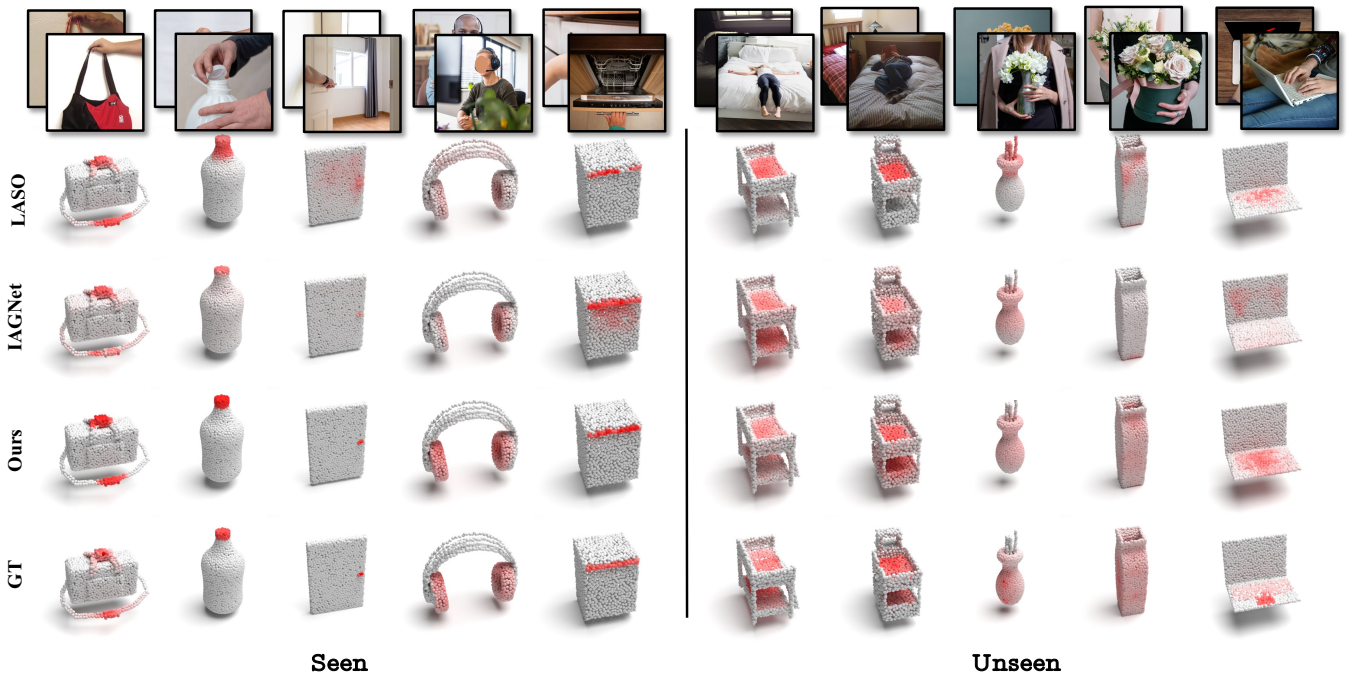


Figure 3: **Affordance Visualization on MIPA dataset.** Compared with LASO (Li et al. 2024c) and IAGNet (Yang et al. 2023), the proposed MIFAG achieves more accurate results in both seen and unseen settings.

	IAM	ADM	AUC ↑	aIOU ↑	SIM ↑	MAE ↓
Seen			82.97	16.88	0.519	0.107
	✓		84.80	20.28	0.555	0.092
		✓	84.57	17.35	0.536	0.117
	✓	✓	**85.10**	**20.50**	**0.568**	**0.091**
Unseen			65.14	4.62	0.311	0.145
	✓		70.34	4.96	0.312	0.150
		✓	69.91	5.21	0.300	0.141
	✓	✓	**71.13**	**5.23**	**0.315**	**0.136**

Table 2: **Ablation of IAM and ADM.** We investigate the improvement of IAM and ADM on the model performance based on the baseline.

	Metrics	Number of Images				
		1	2	3	4	5
Seen	AUC ↑	83.33	84.15	83.68	83.78	**85.47**
	aIOU ↑	19.32	20.16	20.08	18.71	**20.35**
	SIM ↑	0.533	0.550	0.556	0.550	**0.559**
	MAE ↓	0.093	0.093	0.094	0.094	**0.090**
Unseen	AUC ↑	71.05	70.00	70.68	70.63	**71.32**
	aIOU ↑	4.72	4.15	**5.15**	4.15	4.49
	SIM ↑	0.314	0.294	0.301	**0.318**	**0.318**
	MAE ↓	0.146	0.151	0.158	0.154	**0.128**

Table 3: **Ablation of the image numbers** on the **MIPA** dataset.

4.4 Ablation Study

Effectiveness of IAM and ADM. Table 2 reports the effect of IAM and ADM in both **seen** and **unseen** settings of the **MIPA** dataset. Our full model outperforms in all metrics, with higher **AUC** scores of 85.10 and 71.13 compared the model without ADM (84.80 and 70.34) and IAM (84.57 and 69.91) in the **seen** and **unseen** settings, respectively. Notably, the IAM significantly enhances performance, while the ADM has a smaller impact due to its reliance on IAM and feature alignment. Consequently, the best results across all metrics, for both seen and unseen settings, are achieved when IAM and ADM are combined, as shown in the last row of Table 2.

Effect of the image numbers. Table 3 shows the performance of MIFAG with the number of image inputs ranging from 1 to 5. The experimental results indicate a notable performance enhancement in the seen setting with an increasing number of image inputs, as evidenced by improvements in **AUC** (from 83.33 to 85.47) and **aIOU** (from 19.32 to 20.35). Nevertheless, an overabundance of image inputs can lead to a decline in affordance prediction performance in the unseen setting, as reflected in the decrease in **aIOU** from 4.72 to 4.49. This suggests that a limited number of multi-images with significant variations may interfere with the alignment process. In contrast, a sufficient number of images enables the model to distill generalizable and invariant features, resulting in enhanced performance.

Ablation of iterative layer numbers. Table 4 evaluates the invariant affordance feature extraction capability of IAM with different iterative layer numbers. Intuitively, increasing the number of iterative layers can enhance the ability to extract invariant information. However, the optimal layer

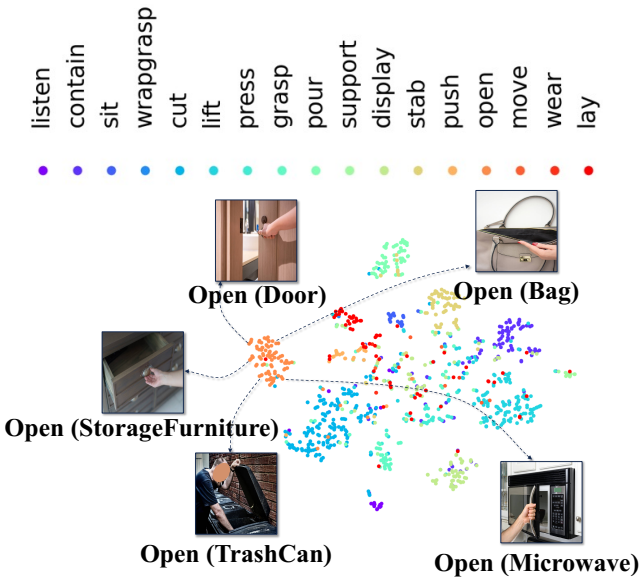


Figure 4: **t-SNE visualization of affordance queries.** Tokens corresponding to the same operation across different objects cluster in the same region.

settings for the seen and unseen scenarios are four and six, respectively, achieving the highest **AUC** (85.10 and 71.13) and **SIM** (0.568 and 0.315). This discrepancy results from a sharp increase in similarity loss as the number of layers grows, which can cause training instability and performance degradation.

	Metrics	Number of Layers					
		1	2	3	4	5	6
Seen	AUC ↑	84.42	84.35	84.56	**85.10**	83.92	83.37
	aIOU ↑	19.83	20.44	**20.70**	20.50	19.32	19.56
	SIM ↑	0.556	0.556	0.560	**0.568**	0.545	0.554
	MAE ↓	0.093	0.093	0.092	**0.091**	0.099	0.094
Unseen	AUC ↑	70.60	69.06	70.87	70.30	70.94	**71.13**
	aIOU ↑	4.70	4.47	4.45	4.72	4.93	**5.23**
	SIM ↑	0.309	0.313	0.313	0.311	0.309	**0.315**
	MAE ↓	**0.131**	0.160	0.164	0.143	0.153	0.136

Table 4: **Ablation of iterative layer numbers** on the **MIPA** dataset.

4.5 Real-World Evaluation

To evaluate the zero-shot generalization ability of our method in real-world scenarios, we test it on real scenes, as shown in Figure 5.

Specifically, we use an iPhone 15 Pro equipped with LiDAR to scan real-world objects and generate their point clouds. These point clouds, along with captured human-object interaction reference images, are then fed into our trained model to generate affordance predictions on the point clouds. We conduct the real-world evaluation under two settings: `seen` and `unseen`, in a manner consistent with previous experiments. The "bed" and the "chair" are objects that are present in the dataset, while the "sofa" is not included in the training dataset. Notably, objects in both settings are new to our trained model, as their point clouds are built from scratch by ourselves. The robust visualization results demonstrate the effective generalization of our method.

Figure 5: **Real-World Visualization. Left:** Original 3D point clouds scanned by an iPhone 15 Pro. **Middle:** Reference images. **Right:** Affordance prediction results on the scanned point cloud.

5 Conclusion and Limitations

In this work, we propose the **M**ulti-**I**mage Guided Invariant-**F**eature-Aware 3D **A**ffordance **G**rounding (**MIFAG**) framework. Our approach gradually extracts affordance knowledge from multiple human-object reference images and effectively integrates this invariant knowledge with point cloud representations to achieve accurate affordance prediction. Moreover, we construct the **M**ulti-**I**mage and **P**oint **A**ffordance (**MIPA**) benchmark to advance research in understanding the affordances of 3D objects. Extensive experiments are conducted on the MIPA dataset and our method outperforms previous state-of-the-art methods.

Limitations Despite the demonstrated effectiveness of MIFAG, it is important to acknowledge certain limitations. Specifically, the 3D objects in our dataset are still relatively simple compared to real-world scenarios. Moreover, our work primarily focuses on visual understanding of affordances, without accounting for manipulation, *e.g.*, the size of the manipulator, the direction of the manipulation action, *etc.* In future work, we aim to bring our approach closer to real-world embodied manipulation. Further discussions and analysis are provided in the supplementary materials.

Acknowledgments

This work is supported by the Shanghai AI Laboratory, the National Key R&D Program of China (2022ZD0160101), the National Natural Science Foundation of China (62376222), and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

Aiello, E.; Valsesia, D.; and Magli, E. 2022. Cross-modal learning for image-guided point cloud shape completion. *Advances in Neural Information Processing Systems*, 35: 37349–37362.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, 1017–1025.

Chen, C.; Cong, Y.; and Kan, Z. 2024. WorldAfford: Affordance Grounding based on Natural Language Instructions. *arXiv preprint arXiv:2405.12461*.

Chen, H.; Wei, Z.; Xu, Y.; Wei, M.; and Wang, J. 2022. ImLoveNet: Misaligned Image-supported Registration Network for Low-overlap Point Cloud Pairs. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393379.

Chen, J.; Gao, D.; Lin, K. Q.; and Shou, M. Z. 2023. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6799–6808.

Chen, S.; Ma, Y.; Qiao, Y.; and Wang, Y. 2024. M-bev: Masked bev perception for robust autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1183–1191.

Cheng, K.; Wu, R.; Shen, Y.; Ning, C.; Zhan, G.; and Dong, H. 2023. Learning Environment-Aware Affordance for 3D Articulated Object Manipulation under Occlusions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dai, Q.; Zhu, Y.; Geng, Y.; Ruan, C.; Zhang, J.; and Wang, H. 2023. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1757–1763. IEEE.

Delitzas, A.; Takmaz, A.; Tombari, F.; Sumner, R.; Pollefeys, M.; and Engelmann, F. 2024. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14531–14542.

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

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

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

Gao, X.; Zhang, P.; Qu, D.; Wang, D.; Wang, Z.; Ding, Y.; Zhao, B.; and Li, X. 2024. Learning 2D Invariant Affordance Knowledge for 3D Affordance Grounding. arXiv:2408.13024.

Geng, H.; Li, Z.; Geng, Y.; Chen, J.; Dong, H.; and Wang, H. 2023a. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2978–2988.

Geng, H.; Xu, H.; Zhao, C.; Xu, C.; Yi, L.; Huang, S.; and Wang, H. 2023b. 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)*, 7081–7091.

Guo, Q.; Dong, Y.; Tian, L.; Kang, Z.; Zhang, Y.; and Wang, S. 2024. BANER: Boundary-Aware LLMs for Few-Shot Named Entity Recognition. *arXiv preprint arXiv:2412.02228*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, S.; Chang, H.; Liu, Y.; Zhu, Y.; Dong, H.; Gao, P.; Boularias, A.; and Li, H. 2024. A3VLM: Actionable Articulation-Aware Vision Language Model. *arXiv preprint arXiv:2406.07549*.

Huang, S.; Jiang, Z.; Dong, H.; Qiao, Y.; Gao, P.; and Li, H. 2023. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*.

Jaritz, M.; Gu, J.; and Su, H. 2019. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Li, G.; Sun, D.; Sevilla-Lara, L.; and Jampani, V. 2024a. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3086–3096.

Li, X.; Sun, P.; Liu, Y.; Duan, L.; and Li, W. 2024b. Simultaneous Detection and Interaction Reasoning for Object-Centric Action Recognition. *arXiv preprint arXiv:2404.11903*.

Li, Y.; Zhao, N.; Xiao, J.; Feng, C.; Wang, X.; and Chua, T.-s. 2024c. LASO: Language-guided Affordance Segmentation on 3D Object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14251–14260.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Y.; Liu, Z.; Zhai, Y.; Li, W.; Doerman, D.; and Yuan, J. 2024. Stat: Towards generalizable temporal action localization. *arXiv preprint arXiv:2404.13311*.

Lobo, J. M.; Jiménez-Valverde, A.; and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2): 145–151.

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

Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; and Tao, D. 2022. Learning Affordance Grounding from Exocentric Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2252–2261.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Mo, K.; Guibas, L. J.; Mukadam, M.; Gupta, A.; and Tulsiani, S. 2021. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6813–6823.

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019. 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)*.

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

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qian, S.; Chen, W.; Bai, M.; Zhou, X.; Tu, Z.; and Li, L. E. 2024. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7587–7597.

Qu, D.; Chen, Q.; Zhang, P.; Gao, X.; Zhao, B.; Wang, D.; and Li, X. 2024. LiveScene: Language Embedding Interactive Radiance Fields for Physical Scene Rendering and Control. *arXiv preprint arXiv:2406.16038*.

Rahman, M. A.; and Wang, Y. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, 234–244. Springer.

Swain, M. J.; and Ballard, D. H. 1991. Color indexing. *International journal of computer vision*, 7(1): 11–32.

Tan, X.; Chen, X.; Zhang, G.; Ding, J.; and Lan, X. 2021. Mbdf-net: Multi-branch deep fusion network for 3d object detection. In *Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data*, 9–17.

Willmott, C. J.; and Matsuura, K. 2005. 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.

Wu, J.; Antonova, R.; Kan, A.; Lepert, M.; Zeng, A.; Song, S.; Bohg, J.; Rusinkiewicz, S.; and Funkhouser, T. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8): 1087–1102.

Wu, R.; Zhao, Y.; Mo, K.; Guo, Z.; Wang, Y.; Wu, T.; Fan, Q.; Chen, X.; Guibas, L.; and Dong, H. 2021. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *arXiv preprint arXiv:2106.14440*.

Xu, C.; Chen, Y.; Wang, H.; Zhu, S.-C.; Zhu, Y.; and Huang, S. 2022a. PartAfford: Part-level Affordance Discovery from 3D Objects. *arXiv preprint arXiv:2202.13519*.

Xu, D.; Anguelov, D.; and Jain, A. 2018. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 244–253.

Xu, X.; Dong, S.; Ding, L.; Wang, J.; Xu, T.; and Li, J. 2022b. FusionRCNN: LiDAR-Camera Fusion for Two-stage 3D Object Detection. *arXiv preprint arXiv:2209.10733*.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; Luo, J.; and Zha, Z.-J. 2023. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10905–10915.

Yu, G.; Li, Y.; Guo, X.; Wang, D.; Liu, Z.; Wang, S.; and Yang, T. 2024a. LiNo: Advancing Recursive Residual Decomposition of Linear and Nonlinear Patterns for Robust Time Series Forecasting. *arXiv preprint arXiv:2410.17159*.

Yu, G.; Zou, J.; Hu, X.; Aviles-Rivero, A. I.; Qin, J.; and Wang, S. 2024b. Revitalizing Multivariate Time Series Forecasting: Learnable Decomposition with Inter-Series Dependencies and Intra-Series Variations Modeling. In *Forty-first International Conference on Machine Learning*.

Zhang, L.; Li, Y.; Zhou, X.; Zhao, X.; and Gu, S. 2024. Transcending the Limit of Local Window: Advanced Super-Resolution Transformer with Adaptive Token Dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2856–2865.

Zhao, L.; Lu, J.; and Zhou, J. 2021. Similarity-Aware Fusion Network for 3D Semantic Segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1585–1592. IEEE.

Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; and Tan, M. 2021. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16280–16290.