GEAL: Generalizable 3D Affordance Learning with Cross-Modal Consistency

Dongyue Lu Lingdong Kong Tianxin Huang Gim Hee Lee
National University of Singapore
dongyue.lu@u.nus.edu lingdong@comp.nus.edu.sg
{huangtx, gimhee.lee}@nus.edu.sg

Project Page: dylanorange.github.io/projects/geal

Abstract

Identifying affordance regions on 3D objects from semantic cues is essential for robotics and human-machine interaction. However, existing 3D affordance learning methods struggle with generalization and robustness due to limited annotated data and a reliance on 3D backbones focused on geometric encoding, which often lack resilience to real-world noise and data corruption. We propose ***GEAL****, a novel framework designed to enhance the generalization and robustness of 3D affordance learning by leveraging large-scale pre-trained 2D models. We employ a dual-branch architecture with Gaussian splatting to establish consistent mappings between 3D point clouds and 2D representations, enabling realistic 2D renderings from sparse point clouds. A granularity-adaptive fusion module and a 2D-3D consistency alignment module further strengthen cross-modal alignment and knowledge transfer, allowing the 3D branch to benefit from the rich semantics and generalization capacity of 2D models. To holistically assess the robustness, we introduce two new corruption-based benchmarks: PIAD-C and LASO-C. Extensive experiments on public datasets and our benchmarks show that GEAL consistently outperforms existing methods across seen and novel object categories, as well as corrupted data, demonstrating robust and adaptable affordance predictions. The code and datasets are publicly available.*

Figure 1. **3D affordance prediction under varied data noises.** Given a textual prompt, previous methods like LASO [32] (right side of each example) exhibit reduced robustness across different corruption types. In contrast, our proposed method, **GEAL** (left side of each example), maintains high accuracy and generalization across these challenging scenarios by effectively transferring knowledge from a large-scale pre-trained 2D foundation model, enhancing robustness and adaptability under diverse conditions.

1. Introduction

3D affordance learning involves identifying interactive regions on objects given semantic cues such as image or textual instruction [7, 10], which is a fundamental competency for intelligent systems [6, 18] to infer how an object can be used or manipulated [27, 42, 70]. This understanding is vital for applications in robotics and human-machine interaction such as action prediction, object manipulation, and autonomous decision-making [5, 12, 13, 16]. For example, a robot equipped with affordance knowledge can intelligently interact with objects in its environment by determining where to grasp a handle or press a button.

Despite its potential, 3D affordance learning still faces significant challenges. Due to limited annotated data, 3D affordance models generally show poorer generalization than their 2D counterparts which benefit from abundant labeled data and large-scale pretraining [28]. Additionally, 3D models often rely on backbones that focus on positional and geometric encoding, limiting their capacity to capture global semantic content and making them vulnerable to noisy or corrupted data from sensor inaccuracies, scene complexity, or processing artifacts in real-world settings [23, 37, 56, 57, 66]. These issues further hinder the robustness and adaptability of existing methods.

In this paper, we introduce a novel framework **GEAL**, which is designed to enhance the generalization and robustness of 3D affordance learning through a dual-branch architecture that leverages the correspondence between 2D and 3D data. GEAL generates realistic 2D renderings directly from sparse 3D data by employing 3D Gaussian splatting (3DGS) [20] to build consistent mappings between 3D point clouds and 2D representations. This approach effectively creates a 2D branch from purely 3D data, which allows us to utilize the generalization capabilities and rich semantic knowledge of large-scale pre-trained 2D foundation models [51, 54] to enhance 3D affordance predictions.

We further introduce a granularity-adaptive fusion module, and a 2D-3D consistency alignment module to ensure robust multi-modal alignment. The granularity-adaptive fusion module dynamically integrates multi-level visual and textual features to address affordance queries at various scales and granularities. The 2D-3D consistency alignment module concurrently establishes reliable cross-modal correspondence with feature embeddings augmented to the Gaussian primitives of 3DGS, fostering effective knowledge transfer across branches, and enhancing the generalization and robustness of the 3D branch by enforcing consistent alignment between 2D and 3D modalities.

In view of the limitation of data scarcity to benchmark the robustness of 3D affordance models, we create two datasets: **PIAD-C**orrupt and **LASO-C**orrupt from existing commonly used affordance datasets [32, 70]. We design these benchmark datasets by incorporating various types of real-world corruptions such as scaling, cropping, *etc.*, to ensure their suitability in evaluating the robustness of 3D affordance models. By contributing these benchmark datasets, we aim to fill a critical gap in the affordance learning community by providing a standard for evaluating the robustness of point cloud-based 3D affordance methods. Fig. 1 shows an example of the text description and the corresponding 3D affordance on 3D point clouds that are corrupted under various noise types.

We validate the generalization and robustness of our GEAL on both standard and corruption-based benchmarks, demonstrating that our approach consistently outperforms recent methods in all scenarios. Our experiments confirm that our GEAL effectively transfers knowledge from seen to unseen data and maintains high performance even under corruption, underscoring the adaptability of our framework across challenging scenarios. In summary, the main contributions of this work can be summarized as follows:

- We propose **GEAL**, a novel approach for generalizable 3D affordance learning. By employing 3DGS, we develop a 2D affordance prediction branch for 3D point clouds, harnessing the robust generalization and semantic understanding of pre-trained 2D foundation models.
- We propose granularity-adaptive fusion and 2D-3D consistency alignment to integrate and propagate knowledge across the dual-branch architecture, and enhance the generalizability of the 3D branch using 2D knowledge.
- We establish two corruption-based benchmarks: *PIAD-C* and *LASO-C*, to holistically evaluate the robustness of 3D affordance learning under real-world scenarios, contributing a standard to the community for robustness analysis.
- Extensive experiments validate the strong performance of our approach on both mainstream and corruption 3D affordance learning benchmarks, proving its generalization ability and robustness across diverse conditions.

2. Related Work

2D Affordance Learning. Affordances refer to potential actions that objects or environments enable for an observer, based on their properties[5, 9, 25, 48]. Early methods for affordance detection mainly try to identify interaction regions in images and videos [29, 40, 58, 61, 71], though these often lacked precise localization of affordance-relevant object parts. To address this, later research improved affordance localization [4, 10, 27, 35, 39, 41, 42, 49, 70] given demonstration 2D data. Recently, large-scale pre-trained models [2, 54] have aligned visual features with affordance-related textual descriptions, reducing dependence on manual labels and enhancing affordance prediction in new contexts [39, 44, 45, 50]. Building on this, some studies [14, 27, 28] turn to leverage foundation models to generalize affordance detection to novel objects and views.

3D Affordance Learning. Extending affordance detection to 3D space presents challenges due to the need for accurate spatial and depth information. While some studies use 2D data to detect 3D affordance regions [5, 9, 29], they often struggle with precise 3D interaction sites. The availability of large-scale 3D object datasets [11, 34, 46] has driven efforts to map affordances directly onto 3D structures [8, 32, 47, 67, 70], aiming to capture complex spatial relationships. Recent methods [19, 32, 50] leverage 2D visual and language models for open-vocabulary affordance detection, enhancing generalization without fixed label sets. Despite these advancements, achieving robust generalization in 3D remains challenging, as 3D backbones still lack the generalization capabilities of 2D foundation models, thus, our method leverages large-scale 2D foundation models to improve 3D affordance generalization.

Robustness for 3D Affordance Learning. Real-world 3D affordance learning faces inevitable challenges from point cloud corruptions caused by scene complexity, sensor inaccuracies, and processing errors [17, 31, 57, 66]. Ex-

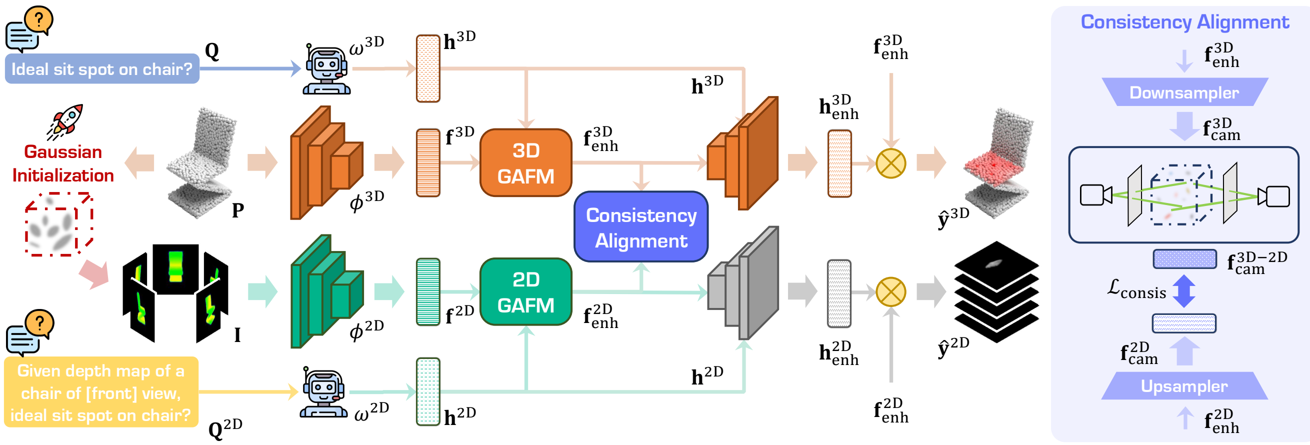

Figure 2. **(Left): Framework Overview.** The proposed **GEAL** consists of two branches: 3D and 2D. The 2D branch is established through 3D Gaussian Splatting to leverage the generalization capabilities of large pre-trained 2D models (*cf*. 3.1). We then perform cross-modality alignment, including **Granularity-Adaptive Visual-Textual Fusion** and **2D-3D Consistency Alignment**, to unify features from different modalities into a shared embedding space (*cf*. 3.2). Finally, we decode generalizable affordance from this embedding space (*cf*. 3.3). **(Right): Architecture of the 2D-3D Consistency Alignment Module.** This module maps features from 2D and 3D modalities into a shared embedding space and enforces consistency alignment to enable effective knowledge transfer across branches.

isting studies aim to improve the robustness against noise and corruption in 3D perception across real-world scenarios [15, 21, 23, 24, 26, 63, 64]. However, affordance learning uniquely requires precise identification of interactive regions under variable and degraded data conditions. To our knowledge, this work is the first to specifically address robustness in 3D affordance learning, providing a targeted approach to enhance reliability across diverse environments.

3. Methodology

In this section, we describe the technical components of our proposed **GEAL** framework. An overview of the full framework is shown in Fig. 2. Given an instruction Q and an object point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$ with N points, GEAL predicts an affordance score $\mathbf{y} \in \mathbb{R}^{N}$, where each value in $\mathbf{y}$ indicates the likelihood that a corresponding point supports the specified functionality. In Sec. 3.1, we employ Gaussian splatting as a cross-modal mapping to bridge the 2D and 3D modalities, establishing a 2D branch to leverage the generalization and robustness of large pre-trained 2D models. Sec. 3.2 details our cross-modal alignment strategy, incorporating both granularity-adaptive visual-text fusion and 2D-3D consistency alignment to unify these modalities in the embedding space. In Sec. 3.3, we outline the decoding process that derives robust and generalizable affordance predictions from the aligned feature space.

3.1. 3D-2D Mapping with Gaussian Splatting

Motivation. Current 3D affordance learning methods suffer from poor generalization due to limited annotated data and exhibit relatively weak robustness owing to limited global semantic capture. In contrast, 2D affordance learning methods [27, 28] leverage 2D foundation models [51, 54] pre-trained on large amounts of data, offering superior generalization and robustness. A 3D-2D mapping to leverage the strengths of 2D foundation models is thus promising. However, a direct projection of 3D point clouds onto 2D planes yields sparse 2D points without semantic and depth information that are not useful for feature extraction with 2D foundation models. To overcome this issue, we adopt 3D Gaussian Splatting [20] which represents 3D scenes as learnable Gaussian primitives for realistic, differentiable and high-speed rendering from arbitrary viewpoints. This approach allows us to synthesize realistic 2D images from sparse point clouds, preserving crucial semantic and depth information for downstream affordance learning tasks. Moreover, 3D Gaussian Splatting offers smoother transitions between points, and preserves occlusions and depth cues for a coherent and accurate scene representation.

Gaussian Initialization. In 3D Gaussian Splatting, each Gaussian primitive is characterized by its 3D position μ represented by a 3D coordinate, a covariance matrix Σ which defines its shape and spread, spherical harmonic parameters c representing its color, and an opacity value α that indicates its transparency. To render 3D Gaussian primitives into 2D image planes, we apply point-based α-blending using a tile-based rasterizer for efficient rendering. The rendered color at each pixel v is calculated as follows:

$$C(v) = \sum\nolimits_{i\in\mathcal{N}} c_i\alpha_i \prod\nolimits_{j=1}^{i-1}(1-\alpha_j), \tag{1}$$

where c_i is the color of the i-th Gaussian, $\mathcal{N}$ represents the Gaussians within the tile, and $\alpha_i = o_i G_i^{\text{2D}}(v)$. o_i is opacity of the i-th Gaussian and $G_i^{\text{2D}}(\cdot)$ represents the function of the i-th Gaussian projected onto 2D. Similarly, a depth map can be rendered as:

$$D(v) = \sum\nolimits_{i\in\mathcal{N}} d_i \alpha_i \prod\nolimits_{j=1}^{i-1}(1-\alpha_j), \tag{2}$$

where d_i denotes the depth of the i-th Gaussian primitive under the provided camera pose.

To ensure that the rendered images accurately reflect the geometry of the input point cloud $\mathbf{P}$, we set the Gaussian mean positions to match the point coordinates, *i.e.* $\boldsymbol{\mu} = \mathbf{P}$. The covariance $\boldsymbol{\Sigma}$ and opacity $\boldsymbol{\alpha}$ are manually adjusted and then kept fixed during training to preserve the original geometry. Using the depth map from Eq. (2) with V camera poses and a predefined color map, we obtain realistic images $\mathbf{I} \in \mathbb{R}^{V\times 3\times H\times W}$ that preserve both semantics and spatial information of the original point cloud, effectively bridging the 3D-2D gap. Treating the affordance score $\mathbf{y} \in [0,1]$ as grayscale color, we assign the color of each Gaussian to match its affordance score, *i.e.* $\mathbf{c} = \mathbf{y}$. We generate 2D affordance masks $\mathbf{y}_{\text{2D}} \in \mathbb{R}^{V\times H\times W}$, where each pixel represents the affordance score of the associated 3D point. This process establishes a coherent mapping from 3D point clouds and affordance scores to their 2D counterparts, using Gaussian splatting to generate realistic, informative 2D representations that enhance affordance learning.

Encoding. Our **GEAL** framework as shown in Fig. 2 comprises a 2D and a 3D branch with backbones $\phi^{\text{2D}}(\cdot)$ and $\phi^{\text{3D}}(\cdot)$, respectively. The 3D branch uses PointNet++ [53] for point cloud feature extraction, while the 2D branch employs DINOV2 [51] for image features. Both networks produce multi-scale features at various granularities. At each scale i, we extract features:

$$\mathbf{f}_i^{\text{3D}} = \phi_i^{\text{3D}}(\mathbf{P}), \quad \mathbf{f}_i^{\text{2D}} = \phi_i^{\text{2D}}(\mathbf{I}), \tag{3}$$

where $\mathbf{f}_i^{\text{3D}} \in \mathbb{R}^{B\times C_i^{\text{3D}}\times N_i^P}$ and $\mathbf{f}_i^{\text{2D}} \in \mathbb{R}^{B\times V\times C^{\text{2D}}\times N^I}$. B is the batch size, V is the number of views, and C_i^{3D} and C^{2D} are feature dimensions. N_i^P is the number of point in scale i, and N^I is image patch length. Note that the 3D spatial resolution N_i^P and C_i^{3D} differ between different scales due to the usage of PointNet++ [53].

We process the input prompt Q using lightweight language models $\omega^{\text{3D}}(\cdot)$ and $\omega^{\text{2D}}(\cdot)$ that share the same architecture. For the 2D input, we modify the prompt to Q^{2D} by adding: "*Given a depth map of a* [object] *in* [view]", constructing **view-dependent prompt** to enhance context understanding. The text embeddings are:

$$\mathbf{h}^{\text{3D}} = \omega^{\text{3D}}(Q), \quad \mathbf{h}^{\text{2D}} = \omega^{\text{2D}}(Q^{\text{2D}}), \tag{4}$$

where $\mathbf{h}^{\text{3D}} \in \mathbb{R}^{B\times C^{\text{txt}}\times L}$ and $\mathbf{h}^{\text{2D}} \in \mathbb{R}^{B\times V\times C^{\text{txt}}\times L}$, with L as the sequence length.

3.2. Cross-Modal Consistency Alignment

Since point cloud, image, and text features are embedded in distinct spaces, we design alignment modules to map these multi-modal features into a shared embedding space. First, we fuse visual features at varying granularities with textual features through Granularity-Adaptive Visual-Textual Fusion, supporting affordance learning conditioned on instructions across different scales. Subsequently, we propagate knowledge from the 2D to 3D branch via a 2D-3D Consistency Alignment by enforcing consistency between 2D and 3D features.

Granularity-Adaptive Visual-Textual Fusion. Both 2D and 3D backbones capture different levels of granularity, with lower layers focusing on fine details and higher layers providing broader context. Since affordances can span multiple object parts, leveraging features at various granularities is advantageous. To achieve this, we introduce **G**ranularity-**A**daptive **F**usion **M**odule **(GAFM)**, which integrates multi-granularity visual features with textual cues via *Flexible Granularity Feature Aggregation* and *Text-Conditioned Visual Alignment*. These mechanisms allow the model to adaptively fuse features across different granularities, enhancing affordance prediction in response to specific instructions. An illustration of the Granularity-Adaptive Fusion Module is provided in Fig. 3.

Flexible Granularity Feature Aggregation. This mechanism aims to fuse visual features from different granularities. Taking the 2D branch as an example, we concatenate feature maps from the last m levels, forming an input tensor $\mathbf{f}_{\text{con}}^{\text{2D}} \in \mathbb{R}^{B\times V\times(m\times C^{\text{2D}})\times N^I}$. Inspired by previous works [28, 72], we then compute adaptive soft weights to regulate the contribution of each feature level, enabling the model to adapt to varying levels of detail. These weights are computed via a gating function with learned noise, introducing perturbations that enhance adaptability:

$$\mathbf{W} = \text{Softmax}\left(\mathbf{f}_{\text{con}}^{\text{2D}} \cdot \mathbf{W}_g + \sigma\cdot\epsilon\right), \tag{5}$$

where $\mathbf{W}_g \in \mathbb{R}^{(m\times C^{\text{2D}})\times m}$ is a trainable weight matrix, $\mathbf{W} \in \mathbb{R}^{B\times m}$ represents the concatenation of weights w_i for each feature level, σ is a learned standard deviation controlling the noise scale, and $\epsilon \sim \mathcal{N}(0,1)$ is Gaussian noise. These weights balance the influence of each feature level, enabling affordance reasoning across different granularities.

The fused feature map is then obtained by applying the adaptive weights to features from each level:

$$\mathbf{f}^{\text{2D}} = \sum\nolimits_{i=1}^{m} w_i \odot \mathbf{f}_i^{\text{2D}}, \tag{6}$$

where $\odot$ denotes element-wise multiplication and $w_i \in \mathbf{W}$. This adaptive aggregation yields a robust feature representation across varying conditions to enhance the generalization ability of the model.

Text-Conditioned Visual Alignment. This module is proposed to integrate visual features with the textual instruction. we follow [32, 62] to feed $\mathbf{f}^{\text{2D}}$ and $\mathbf{h}^{\text{2D}}$ into a transformer block. We first enhance the textual features $\mathbf{h}^{\text{2D}}$ with

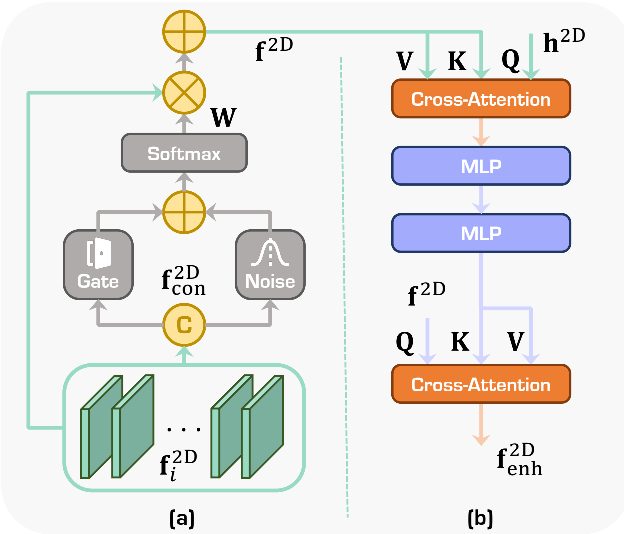

Figure 3. Illustration of the **Granularity-Adaptive Fusion Module**, it consists of a Flexible Granularity Feature Aggregation mechanism (a) and a Text-Conditioned Visual Alignment mechanism (b), we take the 2D branch as an example.

visual features $\mathbf{f}^{2D}$ through cross-attention, followed by refinement with two multilayer perceptrons (MLPs). We then acquire the visual features $\mathbf{f}^{2D}_{enh} \in \mathbb{R}^{B\times C^{txt}\times N^I}$ by querying the refined textual features with cross-attention. We thus ensure that the 2D visual features maintain their spatial structure while embedding the question-relevant information.

In the 3D branch, we align textual features with multi-granularity visual features in a similar manner. However, due to varying spatial resolutions and feature dimensions across scales in PointNet++ [53], directly concatenating all scales' features and computing soft weights as in Eq. (5) is not feasible. To address this, we first apply Text-Conditioned Visual Alignment to the 3D visual features at each scale, then upsample them to a uniform resolution. Finally, we perform Flexible Granularity Feature Aggregation on these upsampled features to produce the aggregated visual representation.

2D-3D Consistency Alignment. 2D features retain rich semantic context and strong generalization via the pre-trained backbone [51], while 3D features preserve geometric and spatial details, compensating for the loss of 2D information caused by self-occlusions. To propagate the knowledge inherently, we introduce **C**onsistency **A**lignment **M**odule **(CAM)** to ensure mutual alignment and knowledge transfer from 2D to 3D spaces.

Specifically, as shown in right part of Fig. 2, we map $\mathbf{f}^{3D}_{enh}$ and $\mathbf{f}^{2D}_{enh}$ into a shared embedding space. Given the 2D-3D correspondence, regions in 2D and 3D representations that map to the same spatial areas should exhibit similar feature representations. By enforcing this consistency, we facilitate 2D-3D knowledge propagation to enhance the understanding of affordances across both modalities of the model.

To align 3D and 2D features in the same embedding space, we employ a down-sampler consisting of two Conv1D layers that reduces the feature dimension of $\mathbf{f}^{3D}_{enh}$ to $\mathbf{f}^{3D}_{cam} \in \mathbb{R}^{B\times C^{cam}\times N}$. This processed feature acts as the representation for each point. We then leverage the established 2D-3D mapping using Gaussian splatting to project these point features into 2D. For each Gaussian, we treat its point feature vector as an inherent attribute. The 2D feature at pixel v is then rendered as:

$$\mathbf{F}(v) = \sum_{i\in\mathcal{N}} \mathbf{f}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{7}$$

where $\mathbf{f}_i$ is the feature of the i-th Gaussian, α_i is its opacity, and $\mathbf{F}(v)$ is the resulting semantic feature at pixel v.

Similarly, we map the 2D features into the same embedding space using an up-sampler consisting of three Conv2D layers, which upsamples the spatial resolution of $\mathbf{f}^{2D}_{enh}$ while also reducing its feature dimension to $\mathbf{f}^{2D}_{cam} \in \mathbb{R}^{B\times V\times C^{cam}\times H\times W}$. V is the number of views and H and W is the spatial dimensions. Given the pixel positions from all V number of $H{*}M$ feature maps as M, we can define the 3D-2D projected feature as $\mathbf{f}^{3D-2D}_{cam} = \{\mathbf{F}(v)|v\in M\}$. We then enforce a consistency constraint by minimizing the difference between the aligned 3D-2D features using L_2 loss:

$$\mathcal{L}_{consis} = \mathrm{MSE}(\mathbf{f}^{3D-2D}_{cam},\ \mathbf{f}^{2D}_{cam}). \tag{8}$$

This consistency loss $\mathcal{L}_{consis}$ encourages the model to maintain similar representations in both 2D and 3D spaces, effectively aligning affordance knowledge across domains. This alignment supports 2D-3D knowledge propagation, ensuring that the information learned in the 2D domain benefits the 3D features.

3.3. Decoding Generalizable Affordance

The affordance scores is decoded under the condition of affordance instructions. Through transformer decoder, the textual features attend to enhanced visual features, focusing the model on specific object parts for accurate predictions.

Our decoder architecture is shared across both 2D and 3D branches. In the 2D branch, textual features $\mathbf{h}^{2D}$ and enhanced visual features $\mathbf{f}^{2D}_{enh}$ are processed through a 3-layer transformer decoder. Here, $\mathbf{h}^{2D}$ serves as the query, and $\mathbf{f}^{2D}_{enh}$ acts as key and value, yielding updated textual features $\mathbf{h}^{2D}_{enh}$. Each layer comprises self-attention to refine textual relationships and cross-attention to guide focus toward relevant visual regions.

After the transformer decoder, these enhanced textual features serve as dynamic kernels to predict affordance scores from visual features. The final affordance prediction $\hat{\mathbf{y}}^{2D}$ is obtained by multiplying $\mathbf{h}^{2D}_{enh}$ with $\mathbf{f}^{2D}_{enh}$, followed by a sigmoid activation:

$$\hat{\mathbf{y}}^{2D} = \mathrm{sigmoid}\left(\mathbf{h}^{2D}_{enh}\cdot\mathbf{f}^{2D}_{enh}\right), \tag{9}$$

where $\hat{\mathbf{y}}^{2D} \in \mathbb{R}^N$ denotes affordance scores.

Training. We employ a combination of Binary Cross-Entropy (BCE) and Dice loss to guide the affordance score prediction in each branch, addressing both class imbalance and segmentation accuracy. For the 2D branch, the loss function is:

$$\mathcal{L}^{2D} = \mathcal{L}^{2D}_{BCE} + \mathcal{L}^{2D}_{Dice}, \tag{10}$$

where $\mathcal{L}^{2D}_{BCE}$ minimizes discrepancies between predicted and true affordance scores, and $\mathcal{L}^{2D}_{Dice}$ improves the overlap between predicted and ground truth regions by maximizing intersection over union.

We adopt a two-stage training approach. We train the 2D branch in the first stage, optimizing it for robust feature extraction and affordance decoding. Except for the CAM (Conisistency Alignment Module), all layers in the 2D branch are frozen in the second stage training. This approach allows the 3D branch to leverage fixed 2D features while adapting to 3D-specific characteristics. Consequently, the loss function for the 3D branch becomes:

$$\mathcal{L}^{3D} = \mathcal{L}^{3D}_{BCE} + \mathcal{L}^{3D}_{Dice} + \mathcal{L}_{consis}. \tag{11}$$

During inference, only the 3D branch is used, ensuring efficient and lightweight affordance prediction.

3.4. Corrupt Data Benchmark

To facilitate robust 3D affordance learning across diverse real-world scenarios, we establish two 3D affordance robustness benchmarks: **PIAD-C** and **LASO-C** based on the test sets of the commonly used datasets **PIAD** and **LASO** following [56]. We apply seven types of corruptions – [1]*Add Global*, [2]*Add Local*, [3]*Drop Global*, [4]*Drop Local*, [5]*Rotate*, [6]*Scale*, and [7]*Jitter* – each with five severity levels. This results in a total of 4,890 object-affordance pairings, comprising 17 affordance categories and 23 object categories with 2,047 distinct object shapes. Due to space limits, more details are provided in the supplementary material.

4. Experiments

4.1. Experimental Settings

Implementation Details. Our model is implemented using PyTorch [52] and is trained using the Adam optimizer [22] with an initial learning rate of 1×10^{-4} for 50 epochs on a single NVIDIA A5000 GPU (with 24 GB memory) with a batch size of 12. A step learning rate scheduler aids convergence. Additionally, the 2D backbone DINOV2 [51] remains frozen during training, while the language model RoBERTa [36] has been fine-tuned.

Datasets. We conduct experiments on **LASO** [32] and **PIAD** [70], both providing paired affordance and point cloud data. **LASO** contains 19,751 point cloud-question pairs over 8,434 objects (23 classes, 17 affordance categories) with *Seen* and *Unseen* splits to test generalization to novel affordance-object pairs. **PIAD** comprises 7,012 point clouds of the same categories, but some objects are entirely unseen during training, challenging the model's generalization to novel objects. Since PIAD lacks language annotations, we reuse LASO by randomly assigning questions to each affordance-object pair.

Metrics. We use four metrics to assess performance: **AUC** [38] measures the ability to rank points correctly; **aIoU** [55] quantifies the overlap between predictions and ground truth; **SIM** [59] assesses the similarity by summing minimum values at each point; and **MAE** [65] calculates the average ab-

Table 1. The overall results of all comparative methods on **PIAD** [70]. **Seen** and **Unseen** are two partitions of the dataset. AUC and aIOU are shown in percentage. The **best** and <u>2nd best</u> scores from each metric are highlighted in **bold** and <u>underlined</u>, respectively.

Type	Method	aIoU ↑	AUC ↑	SIM ↑	MAE ↓
Seen	MBDF [60]	9.3	74.9	0.415	0.143
	PMF [73]	10.1	75.1	0.425	0.141
	FRCNN [69]	12.0	76.1	0.429	0.136
	ILN [3]	11.5	75.8	0.427	0.137
	PFusion [68]	12.3	77.5	0.432	0.135
	XMF [1]	12.9	78.2	0.441	0.127
	IAGNet [70]	<u>20.5</u>	<u>84.9</u>	0.545	0.098
	LASO [32]	19.7	84.2	<u>0.590</u>	<u>0.096</u>
	GEAL (Ours)	**22.5**	**85.0**	**0.600**	**0.092**
Unseen	MBDF [60]	4.2	58.2	0.325	0.213
	PMF [73]	4.7	60.3	0.330	0.211
	FRCNN [69]	5.1	61.9	0.332	0.195
	ILN [3]	4.7	59.7	0.325	0.207
	PFusion [68]	5.3	61.9	0.330	0.193
	XMF [1]	5.7	62.6	0.342	0.188
	IAGNet [70]	<u>8.0</u>	<u>71.8</u>	0.352	0.127
	LASO [32]	<u>8.0</u>	69.2	<u>0.386</u>	<u>0.118</u>
	GEAL (Ours)	**8.7**	**72.5**	**0.390**	**0.102**

Table 2. The overall results of all comparative methods on the **LASO** dataset [32]. **Seen** and **Unseen** are two partitions of the dataset. Results marked with * denote our reproduced results, following the data split reported in LASO [32]. AUC and aIOU are shown in percentage. The **best** and <u>2nd best</u> scores from each metric are highlighted in **bold** and <u>underlined</u>, respectively.

Type	Method	aIoU ↑	AUC ↑	SIM ↑	MAE ↓
Seen	ReferTrans [30]	13.7	79.8	0.497	0.124
	ReLA [33]	15.2	78.9	0.532	0.118
	3D-SPS [43]	11.4	76.2	0.433	0.138
	IAGNet [70]	17.8	82.3	0.561	0.109
	LASO [32]	<u>20.8</u>	**87.3**	<u>0.629</u>	<u>0.093</u>
	LASO* [32]	19.7	85.2	0.600	0.097
	GEAL (Ours)	**22.0**	<u>86.7</u>	**0.634**	**0.092**
Unseen	ReferTrans [30]	10.2	69.1	0.432	0.145
	ReLA [33]	10.7	69.7	0.429	0.144
	3D-SPS [43]	7.9	68.8	0.402	0.158
	IAGNet [70]	12.9	77.8	0.443	0.129
	LASO [32]	14.6	80.2	0.507	0.119
	LASO* [32]	<u>15.6</u>	<u>79.9</u>	<u>0.549</u>	<u>0.108</u>
	GEAL (Ours)	**16.7**	**80.9**	**0.567**	**0.106**

Table 3. Comparison of different methods under various corruption settings on the proposed **PIAD-C** benchmark, evaluated on the **Seen** partition. **Drop-L** denotes local drop, and **Drop-G** denotes global drop; similarly, **Add-L** and **Add-G** refer to local and global addition, respectively. AUC and aIOU are reported as percentages. For each metric, the **best** scores are highlighted in **bold**.

Type	aIOU ↑		AUC ↑		SIM ↑		MAE ↓	
	LASO	GEAL	LASO	GEAL	LASO	GEAL	LASO	GEAL
Scale	17.6	**19.7**	82.1	**82.5**	0.554	**0.562**	0.100	**0.097**
Jitter	14.7	**17.0**	80.3	**80.6**	0.501	**0.505**	0.103	**0.099**
Rotate	16.7	**19.0**	82.2	**82.4**	0.542	**0.550**	0.101	**0.097**
Drop-L	10.6	**12.4**	77.0	**77.2**	0.470	**0.474**	0.112	**0.111**
Drop-G	18.7	**21.1**	83.1	**83.7**	0.545	**0.559**	0.097	**0.094**
Add-L	15.7	**18.5**	81.0	**81.1**	0.525	**0.536**	0.100	**0.095**
Add-G	13.4	**16.1**	76.9	**77.4**	0.506	**0.513**	0.101	**0.098**

Table 4. Comparison of different methods under various corruption settings on the proposed **LASO-C** benchmark, evaluated on the **Seen** partition. **Drop-L** denotes local drop, and **Drop-G** denotes global drop; similarly, **Add-L** and **Add-G** refer to local and global addition, respectively. AUC and aIOU are reported as percentages. For each metric, the **best** scores are highlighted in **bold**.

Type	aIOU ↑		AUC ↑		SIM ↑		MAE ↓	
	LASO	GEAL	LASO	GEAL	LASO	GEAL	LASO	GEAL
Scale	18.7	**21.0**	84.6	**85.3**	0.590	**0.600**	0.103	**0.100**
Jitter	15.4	**17.8**	81.3	**81.9**	0.516	**0.517**	0.107	**0.106**
Rotate	17.8	**19.8**	83.6	**84.3**	0.572	**0.573**	0.101	**0.100**
Drop-L	12.6	**13.3**	79.3	**80.0**	0.466	**0.484**	0.122	**0.110**
Drop-G	18.4	**20.9**	83.5	**85.2**	0.565	**0.567**	0.099	**0.095**
Add-L	17.6	**20.2**	82.7	**84.4**	0.566	**0.572**	0.103	**0.100**
Add-G	16.7	**19.0**	81.1	**83.4**	0.549	**0.575**	0.108	**0.097**

solute difference between predictions and ground truth.

Baselines. We primarily compare our method with state-of-the-art approaches LASO [32] and IAGNet [70]. On **PIAD**, we evaluate against IAGNet and several image-point cloud cross-modal baselines, retraining LASO for comparison. On **LASO**, we compare with the original LASO method and other methods utilizing vision-language models for cross-modal alignment. To adapt IAGNet to LASO, its image backbone is replaced with a language model [32], keeping the rest of the architecture intact. Since the *Unseen* data split of LASO is not publicly available, we reproduce it based on the descriptions in their paper and report our results accordingly. Further experimental details are provided in the supplementary material.

4.2. Comparisons to State-of-the-Art Methods

Seen Categories: In Tab. 1 and Tab. 2, we present the performance of our **GEAL** compared to state-of-the-art approaches on the PIAD and LASO datasets under the *Seen* category setting. On the PIAD dataset, GEAL achieves the highest scores across all evaluation metrics, surpassing the previous best method, IAGNet [70]. Similarly, on the LASO dataset, GEAL outperforms LASO [32] on the majority of metrics. These results highlight the effectiveness of GEAL in leveraging the rich semantic understanding from pre-trained 2D models through Gaussian splatting. The granularity-adaptive fusion and 2D-3D consistency alignment modules enable multi-granularity feature fusion and efficient knowledge transfer between the 2D and 3D modalities, enhancing the ability of the model to accurately predict affordance regions on the seen categories.

Table 5. Ablation study on the impact of different components in **GEAL** on the **PIAD** dataset [70]. **Seen** and **Unseen** are two partitions of the dataset. **2D** denotes the use of the 2D baseline with a weighted sum mapping back to 3D. **3D** represents the 3D baseline. **CAM** is the consistency alignment module. **GAFM** is the granularity-adaptive fusion module. AUC and aIoU are shown in percentage. The **best** and <u>second best</u> scores from each metric are highlighted in **bold** and <u>underlined</u>, respectively.

Type	2D	3D	CAM	GAFM	aIoU ↑	AUC ↑	SIM ↑	MAE ↓
Seen	✓	✗	✗	✗	19.2	80.5	0.567	0.101
	✗	✓	✗	✗	19.5	83.5	0.585	0.097
	✓	✓	✓	✗	<u>22.0</u>	<u>84.4</u>	<u>0.592</u>	<u>0.094</u>
	✓	✓	✓	✓	**22.5**	**85.0**	**0.600**	**0.092**
Unseen	✓	✗	✗	✗	8.5	70.8	0.357	0.112
	✗	✓	✗	✗	8.0	69.2	<u>0.386</u>	0.118
	✓	✓	✓	✗	<u>8.6</u>	<u>71.2</u>	0.371	<u>0.105</u>
	✓	✓	✓	✓	**8.7**	**72.5**	**0.390**	**0.102**

Table 6. Ablation study on the configuration of the Gaussian Splatting parameters in **GEAL** on the **PIAD** dataset [70]. r denotes the resolution, V is the number of views, and **prompt** indicates whether a view-dependent prompt is used. **Seen** and **Unseen** are two partitions of the dataset. AUC and aIoU are shown in percentage. The **best** and <u>2nd best</u> scores from each metric are highlighted in **bold** and <u>underlined</u>, respectively.

Type	r	V	prompt	aIoU ↑	AUC ↑	SIM ↑	MAE ↓
Seen	112	6	✗	20.2	83.5	0.566	0.112
	112	12	✗	21.4	83.8	0.578	0.105
	112	12	✓	22.5	85.0	0.600	0.092
	112	14	✓	22.5	85.2	0.599	0.092
	224	14	✓	**22.9**	**86.1**	**0.603**	**0.089**
Unseen	112	6	✗	7.0	70.7	0.355	0.108
	112	12	✗	7.5	71.9	0.390	0.106
	112	12	✓	8.7	72.5	0.390	0.102
	112	14	✓	8.9	72.8	0.391	0.098
	224	14	✓	**9.2**	**73.0**	**0.394**	**0.095**

Unseen Categories: The *Unseen* category setting evaluates the generalization ability of the model to novel objects not encountered during training. On both the PIAD (Tab. 1) and LASO (Tab. 2) datasets, **GEAL** consistently outperforms all baselines across metrics. Although the absolute performance values are lower due to the increased difficulty of unseen categories, GEAL maintains a performance edge over the baselines. This demonstrates that GEAL effectively generalizes to unseen categories, a result attributed to the integration of the 2D branch with a pretrained foundation model backbone and the cross-modal consistency alignment between the 2D and 3D branches. Qualitative comparisons with LASO on PIAD are shown in Fig. 4.

Robustness on Corrupt Data: To assess robustness under real-world conditions, we compare **GEAL** with LASO on the proposed PIAD-C and LASO-C benchmarks after train-

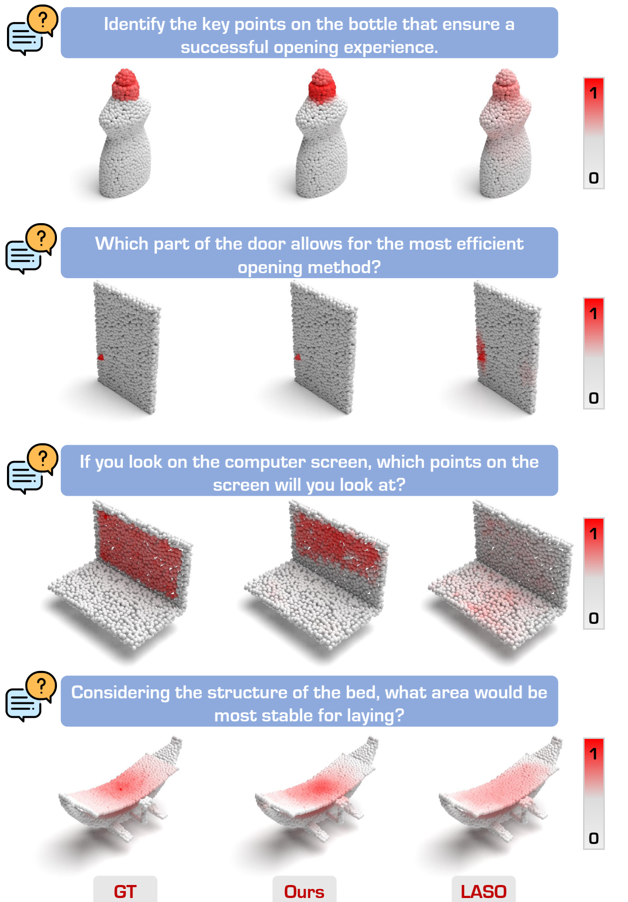

Figure 4. Qualitative comparisons between **GEAL** and LASO [32] on the PIAD [70] dataset. Top two rows display results on *seen* partition, while bottom two rows show results on *unseen* partition. Our method demonstrates strong generalization on both seen and unseen partitions. *cf*. supplementary material for more examples.

ing on clean data. As shown in Tab. 3 and Tab. 4, GEAL consistently outperforms LASO across all corruption types and evaluation metrics. GEAL demonstrates superior resilience under various corruptions, achieving higher AUC and SIM scores while maintaining lower MAE values. This consistent outperformance indicates that the architecture of GEAL effectively mitigates the impact of data degradation. The robustness improvements are attributed to our dual-branch architecture and the 2D-3D consistency alignment module. By leveraging the robustness of pre-trained 2D models and enforcing cross-modal consistency, GEAL maintains high performance even when faced with corrupted or noisy 3D data.

4.3. Ablation Study

Component Analysis. As shown in Tab. 5, we conduct an ablation study on PIAD [70] to evaluate the effectiveness of each component in our proposed **GEAL** framework. We examine the impact of using only the 2D baseline with a weighted sum mapping back to 3D using the inverse process of Eq. (1) (*i.e*., **2D**), only the 3D baseline (*i.e*., **3D**), the consistency alignment module (**CAM**), and the granularity-adaptive fusion module (**GAFM**). Both the 2D and 3D baselines use only the last-layer features from their respective visual backbones to fuse with textual features without considering granularity. The results show that using only the 2D branch or only the 3D branch yields similar baseline performance. Integrating both branches with the consistency alignment module (**CAM**) leads to a noticeable improvement. Finally, our full model incorporating all components, including the granularity-adaptive fusion module (**GAFM**), achieves the best performance on both the *Seen* and *Unseen* sets. This demonstrates the effectiveness of our dual-branch architecture and underscores the importance of granularity-adaptive fusion and cross-modal consistency in enhancing the generalization capability of the model.

Gaussian Splatting Configuration Tuning. As shown in Table 6, we further investigate the impact of different Gaussian splatting configurations on the performance of our model. Specifically, we vary the rendering resolution (r), the number of views (V), and the inclusion of view-dependent prompts. The results indicate that increasing the number of views from 6 to 12 slightly improves performance, suggesting that additional viewpoints provide richer information for affordance learning. Incorporating view-dependent prompts significantly boosts performance, particularly on the *Seen* set, highlighting the importance of semantic guidance in our framework. Increasing the resolution from 112 to 224 yields only marginal gains, indicating that our model is robust to resolution changes and that higher resolutions offer diminishing returns. Balancing effectiveness and efficiency, we opt for a configuration of $r = 112$, $V = 12$, and the use of view-dependent prompts.

5. Conclusion

In conclusion, we present **GEAL**, a framework that significantly improves the generalization and robustness of 3D affordance learning by leveraging large-scale pre-trained 2D foundation models. Through a dual-branch architecture enabled by 3D Gaussian splatting, GEAL maps 3D point clouds to 2D representations, enabling realistic and semantically rich renderings from sparse data. The granularity-adaptive fusion and 2D-3D consistency alignment modules ensure robust cross-modal alignment and effective knowledge transfer, allowing the 3D branch to leverage rich semantic information from the powerful 2D foundation models. Extensive experiments on public datasets and our newly proposed corruption-based benchmarks further show that GEAL consistently outperforms existing methods, demonstrating strong generalization, enhanced adaptability, and robust affordance prediction under a wide range of challenging real-world conditions.

Acknowledgments

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021), and the Tier 2 grant MOE-T2EP20124-0015 from the Singapore Ministry of Education. The authors would like to sincerely thank the Program Chairs, Area Chairs, and Reviewers for the time and efforts devoted during the review process.

References

[1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. In *Advances in Neural Information Processing Systems*, pages 37349–37362, 2022.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[3] 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 Conference Proceedings*, pages 1–9, 2022.

[4] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023.

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

[6] Francisco Cruz, Sven Magg, Cornelius Weber, and Stefan Wermter. Training agents with interactive reinforcement learning and contextual affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):271–284, 2016.

[7] Leiyao Cui, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Yixin Zhu. Strap: Structured object affordance segmentation with point supervision. *arXiv preprint arXiv:2304.08492*, 2023.

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

[9] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *IEEE International Conference on Robotics and Automation*, pages 5882–5889, 2018.

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

[11] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[12] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *IEEE International Conference on Robotics and Automation*, pages 5880–5886, 2023.

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

[14] Denis Hadjivelichkov, Sicelukwanda Zwane, Lourdes Agapito, Marc Peter Deisenroth, and Dimitrios Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pages 550–560. PMLR, 2023.

[15] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, 2024.

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

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[18] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.

[19] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pages 540–562. PMLR, 2023.

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):139–1, 2023.

[21] Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J Kim. Point cloud augmentation with weighted local transformations. In *IEEE/CVF International Conference on Computer Vision*, pages 548–557, 2021.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[23] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[24] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, pages 21298–21342, 2023.

[25] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from

rgb-d videos. *International Journal of Robotics Research*, 32 (8):951–970, 2013.

[26] Dogyoon Lee, Jaeha Lee, Junhyeop Lee, Hyeongmin Lee, Minhyeok Lee, Sungmin Woo, and Sangyoun Lee. Regularization strategy for point cloud via rigidly mixed sample. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15900–15909, 2021.

[27] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

[28] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024.

[29] 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 *IEEE/CVF International Conference on Computer Vision*, pages 12032–12042, 2023.

[30] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *Advances in Neural Information Processing Systems*, pages 19652–19664, 2021.

[31] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your lidar placement optimized for 3d scene understanding? In *Advances in Neural Information Processing Systems*, 2024.

[32] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024.

[33] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023.

[34] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022.

[35] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[36] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[37] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, pages 37193–37229, 2023.

[38] 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.

[39] Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, and Yang Cao. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence*, 4(5):1186–1198, 2023.

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

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

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

[43] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.

[44] 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.

[45] 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.

[46] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[47] 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.

[48] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation*, pages 1374–1381, 2015.

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

[50] Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5692–5698, 2023.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[53] 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.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[55] 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, 2016.

[56] Jiawei Ren, Lingdong Kong, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud robustness under corruptions. *Preprint*, 2022.

[57] Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning*, pages 18559–18575. PMLR, 2022.

[58] 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.

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

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

[61] 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.

[62] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[63] Jie Wang, Lihe Ding, Tingfa Xu, Shaocong Dong, Xinli Xu, Long Bai, and Jianan Li. Sample-adaptive augmentation for point cloud recognition against real-world corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 14330–14339, 2023.

[64] Jie Wang, Tingfa Xu, Lihe Ding, and Jianan Li. Target-guided adversarial point cloud transformer towards recognition against real-world corruptions. *arXiv preprint arXiv:2411.00462*, 2024.

[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.

[66] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

[67] 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.

[68] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018.

[69] Xinli Xu, Shaocong Dong, Lihe Ding, Jie Wang, Tingfa Xu, and Jianan Li. Fusionrcnn: Lidar-camera fusion for two-stage 3d object detection. *Remote Sensing*, 15(7):1839, 2023.

[70] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *IEEE/CVF International Conference on Computer Vision*, pages 10905–10915, 2023.

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

[72] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. *arXiv preprint arXiv:2401.17868*, 2024.

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