One-Shot Open Affordance Learning with Foundation Models

Gen Li[1] Deqing Sun[2] Laura Sevilla-Lara[1] Varun Jampani[3]

[1]University of Edinburgh [2]Google Research [3]Stability AI

Abstract

We introduce One-shot Open Affordance Learning (OOAL), where a model is trained with just one example per base object category, but is expected to identify novel objects and affordances. While vision-language models excel at recognizing novel objects and scenes, they often struggle to understand finer levels of granularity such as affordances. To handle this issue, we conduct a comprehensive analysis of existing foundation models, to explore their inherent understanding of affordances and assess the potential for data-limited affordance learning. We then propose a vision-language framework with simple and effective designs that boost the alignment between visual features and affordance text embeddings. Experiments on two affordance segmentation benchmarks show that the proposed method outperforms state-of-the-art models with less than 1% of the full training data, and exhibits reasonable generalization capability on unseen objects and affordances. Project page: https://reagan1311.github.io/ooal.

1. Introduction

Affordances are the potential "action possibilities" regions of an object [20, 22], which play a pivotal role in various applications, including robotic learning [19, 26, 42], scene understanding [12, 32, 51], and human-object interaction [23, 41]. In particular, affordance is crucial for embodied intelligence, since it facilitates agents' understanding of the associations between objects, actions, and effects in dynamic environments, thus bridging the gap between passive perception and active interaction [14, 38].

Learning to recognize object affordances across a variety of scenarios is challenging, since different objects can vary significantly in appearance, shape, and size, yet have the same functionality. For instance, a chef's knife and a pair of office scissors share common affordances of cutting and holding, but their blades and handles look different.

A large portion of the work [12, 16, 17, 39, 43, 44] has focused on learning a mapping between visual features and affordance labels, utilizing diverse resources as inputs, such as 2D images, RGB-D data, and 3D point clouds. This mapping can be established through a labeled dataset with predefined objects and affordances. However, large-scale affordance datasets are scarce, and most of them have a small number of object categories, making it difficult to apply the learned mapping to novel objects and scenes. To reduce the reliance on costly annotation, some recent studies perform affordance learning from sparse key points [15, 52, 53], videos of humans in action [18, 30, 41], or human-object interaction images [28, 35]. While alleviating the need for dense pixel labeling, these methods still require a large amount of training data. In addition, they often struggle to generalize to unseen objects and cannot identify novel affordances.

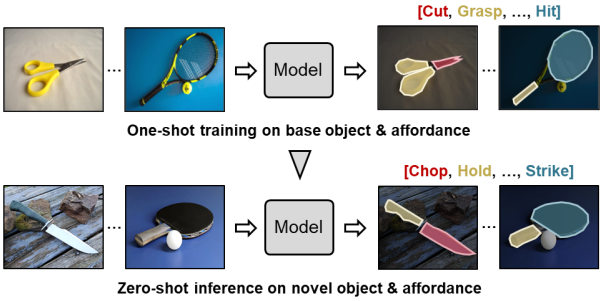

Figure 1. The pipeline of one-shot open affordance learning. It uses one image per base object for training, and can perform zero-shot inference on novel objects and affordances.

To tackle the above limitations, we are interested in learning an affordance model that does not rely on extensive datasets, and can comprehend unseen object and affordance classes. For example, after a model is trained with the knowledge that scissor blades afford cutting, it should generalize to related objects such as knives and axes, inferring that their blades can cut objects too. Moreover, the model should be able to reason about semantically similar vocabularies, *e.g.*, "hold" and "grasp", "cut" and "slice", instead of knowing only predefined affordance categories.

In this paper, we target the extreme case of using merely one example from each base object category and term this

research problem as One-shot Open Affordance Learning (OOAL), where the model is trained with very little data, and is expected to recognize unseen objects and affordances during inference. The illustration of OOAL pipeline is shown in Fig. 1. Compared with the typical affordance learning that requires numerous training samples and can only reason within a closed affordance vocabulary, OOAL alleviates the need of large-scale datasets and broadens the scope of inference.

To this end, we note that foundation Vision-Language Models (VLMs) can be a potential solution, which have recently emerged as powerful tools for a wide array of computer vision tasks. The open vocabulary nature of these VLMs like CLIP [49] that are trained on a large corpus of image-text data enables reasoning of previously unseen objects, scenes, and concepts. However, we observe that these models often fail to understand nuanced vocabularies such as affordances or object parts. One hypothesis is that object parts and affordances appear much less frequently in image captions compared with objects. Therefore, the following question naturally arises: *Can we teach foundation models to comprehend more subtle, fine-grained aspects of objects, such as affordances, with very few examples?* In this way, the generalization capability of foundation models can be inherited with minimum annotation effort.

To achieve this, we first conduct a thorough analysis of several representative foundation models. The objective is to delve into their inherent understanding of affordances, and figure out what visual representation is suitable for data-limited affordance learning. Based on the analysis, we then build a learning architecture and propose several methods, including text prompt learning, multi-layer feature fusion, and a CLS-token-guided transformer decoder, that can facilitate the alignment between visual representation and affordance text embeddings. Lastly, we select a dense prediction task, affordance segmentation, for evaluation and comparison with a variety of state-of-the-art models, where we find that our methods can achieve higher performance with less than 1% of the complete training data.

Overall, our contributions can be summarized as follows: (1) We introduce the problem of OOAL, aiming to develop a robust affordance model that can generalize to novel object and affordance categories without the need of massive training data. (2) We conduct a comprehensive analysis on existing foundation models to explore their potential for OOAL. Following the analysis, we build a learning architecture with vision-language foundation models, and design several methods to improve the alignment between visual features and affordance text labels. (3) We implement extensive experiments with two affordance segmentation datasets to demonstrate the effectiveness of our learning pipeline, and observe noticeable gains over baselines with strong generalization capability.

2. Related Work

Affordance Learning. The term "affordance" is popularized by the psychologist James Gibson, who describes it as the properties of an object or the environment that suggest possible actions or interactions. Building on this, researchers have developed many approaches to acquire affordance information in various ways. In computer vision, initial research [12, 17, 27, 40] has focused on affordance detection using convolutional neural networks. As manual affordance annotations are often costly to acquire, much subsequent research has shifted its focus to weak supervision such as keypoints [15, 52, 53] or image-level labels [35, 41]. Recent work has explored a novel perspective on how to ground affordances from human-object interaction images [28, 35, 63] or human action videos [9, 18, 30, 41]. In robotics, affordance learning enables robots to interact effectively and intelligently with complex and dynamic environments [2, 62]. Specifically, some work [3, 26, 57] utilizes affordance to build relationships between objects, tasks, and manipulations for robotic grasping. Other studies focus on learning affordance from resources that can be deployed on real robots, such as human teleoperated play data [6], image pairs [5], and egocentric video datasets [4].

In contrast to the works above that often require a large amount of training data, we propose the problem of OOAL that aims to perform affordance learning with one sample per base object category, and allows zero-shot inference to handle unseen objects and affordances.

Foundation Models for Affordance Learning. With the rapid development of foundation models such as Large Language Models (LLMs) and Visual Foundation Models (VFMs), many research efforts have explored their utilization in affordance learning or reasoning. Mees *et al*. [36] leverage GPT-3 [7] to break down language instructions into subgoals, and learn a visual affordance model to complete real world long-horizon tasks. Li *et al*. [28] adopt DINO-ViT features to perform affordance grounding by transferring affordance knowledge from human-object interaction images to egocentric views. Huang *et al*. [24] propose a novel pipeline that uses LLMs [46] for affordance reasoning, which interacts with VLMs to produce 3D affordance maps for robotic manipulation. Qian *et al*. [48] develop an approach that employs the rich world knowledge of VLMs to ground affordance, exhibiting powerful generalization ability. Recent studies [37, 50, 55] delve into the integration of affordance and language models for task-oriented grasping, which allows robots to grasp objects in a more appropriate and safe manner.

The closest methods to ours are AffCorrs [21] and OpenAD [45]. AffCorrs utilizes the visual foundation model DINO to find corresponding affordances in a one-shot manner, but relevant objects are explicitly selected as support

images to significantly reduce the difficulty. OpenAD takes advantage of CLIP for open-vocabulary affordance detection in point clouds. It requires a large number of manual annotations, while our work performs affordance learning with merely one example per base object category.

3. One-Shot Open Affordance Learning

3.1. Problem Setting

One-shot Open Affordance Learning (OOAL) aims to learn a model to predict affordance with one example per base object class and can generalize to novel object classes. In this work, we focus on the dense prediction task of affordance segmentation. Specifically, objects are first divided into N_b base classes and N_o novel classes without intersection. The model receives only N_b samples during training, one for each base object category, which is a pair of image $I \in \mathbb{R}^{H\times W\times 3}$ and pixel-wise affordance annotation $M \in \mathbb{R}^{H\times W\times N}$ (N is the number of affordance categories in the dataset). After training, evaluation is performed on the combination of base and novel object categories to measure the generalization ability of the model. Also, affordance labels can be replaced with novel vocabularies that share similar semantics, such as "chop", "slice", and "trim" to represent affordance akin to "cut".

It is worth noting that OOAL is different from one-shot semantic segmentation (OSSS) [54] and one-shot affordance detection (OS-AD) [33]. Both OSSS and OS-AD receive one-shot sample during training. However, the sample keeps changing in each iteration, so the model has access to a large set of image-mask pairs. Additionally, a support image is required at inference to provide prior information. In comparison, OOAL performs one-shot training and zero-shot inference, which poses additional challenges. The model needs to generalize to previously unseen objects, necessitating the ability to understand and recognize semantic relationships between seen and unseen classes with very limited data.

3.2. Analysis of Foundation Models

The field of computer vision has recently witnessed a surge in the prevalence of large foundational models, such as CLIP [49], Segment Anything [25], DINO [8, 47] *etc*. These models exhibit strong zero-shot generalization capabilities for several computer vision tasks, making them seem like a great option to tackle the problem of OOAL. To this end, we perform analysis on several existing foundation models which we split into three parts: ❶ Do current vision-language foundation models and their variants have the ability to detect affordances via affordance/part-based prompting? ❷ Can the features of visual foundation models discriminate affordance regions in images? and ❸ Can these models generalize affordance recognition to novel objects and perform well in the low-shot setting?

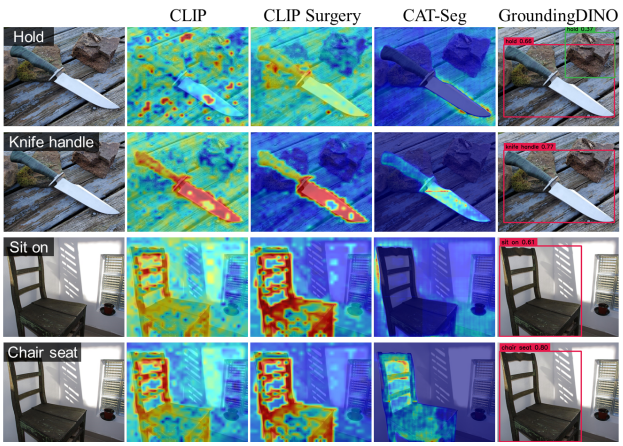

Figure 2. **Analysis of vision-language foundation models** on text-based affordance grounding. The 1st and 3rd rows use affordance texts as input queries, and the 2nd and 4th rows use corresponding object parts as input text queries. Visualizations show that these models have limited ability to recognize fine-grained affordances and object parts.

Driven by question ❶, we select four representative models, *i.e*., the vanilla CLIP, a CLIP-based explainability method CLIP Surgery [29], a state-of-the-art open-vocabulary segmentation method CAT-Seg [11], and an open-vocabulary detection method GroundingDINO [31]. For vanilla CLIP, we employ the method proposed in MaskCLIP [65] that directly extracts dense predictions without fine-tuning. We use the text prompt template of "somewhere to [*affordance*]" to query visual features to find corresponding areas. As illustrated in Fig. 2, we note that most models cannot understand affordance well, except the detection model GroundingDINO, but its predictions mainly focus on the whole object rather than parts. As for dense prediction models, CAT-Seg often recognizes affordance regions as background, and CLIP gives high activation on both foreground and background. In comparison, CLIP Surgery fails to localize the "holding" area for a knife, but manages to associate the phrase "sit on" with a chair. Furthermore, even when the affordance text is replaced with corresponding object parts, predictions from CLIP and GroundingDINO remain biased toward objects, while CLIP Surgery and CAT-Seg tend to activate the wrong parts. This is consistent with recent findings [56, 59] that CLIP has limited part recognition ability.

To answer questions ❷ and ❸, we consider two essential characteristics of a good affordance model in the low-shot setting: (1) Part-aware representation. The visual representation should exhibit awareness of object parts, given that affordance often denotes small and fine-grained regions, *e.g*., a bicycle saddle to sit on or a knife handle to

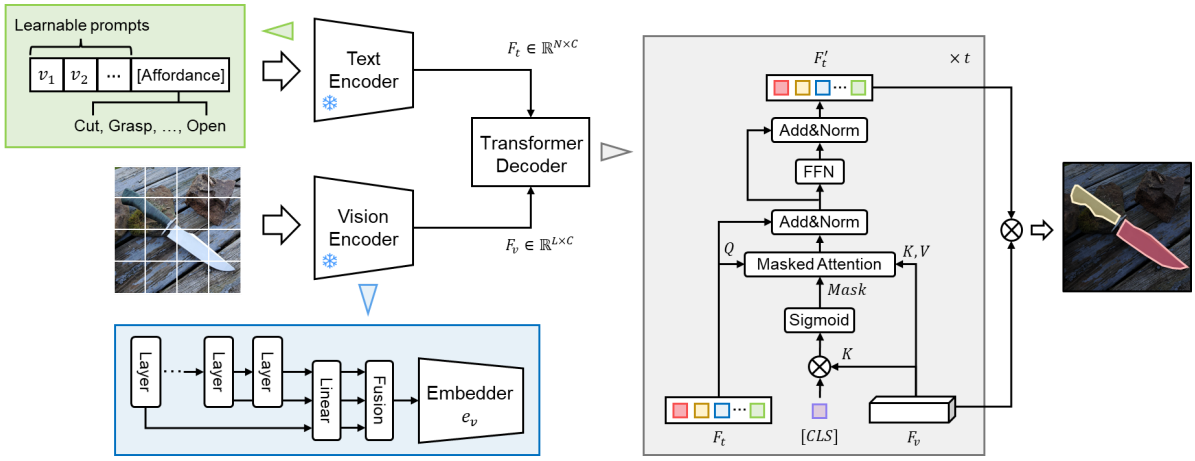

Figure 3. Proposed learning framework for OOAL. Our designs are highlighted in three color blocks, which are text prompt learning, multi-layer feature fusion, and CLS-guided transformer decoder. [CLS] denotes the CLS token of the vision encoder.

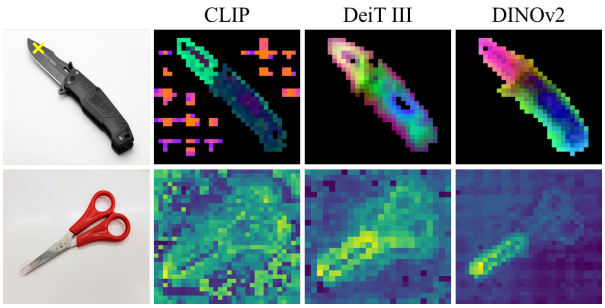

Figure 4. **Analysis of visual foundation models** on affordance learning. Top row: visualizations of PCA components. Bottom row: feature similarity maps between the yellow mark on the knife blade and the image of scissors. Qualitative results show that DINOv2 has clearer part-aware representations and better part-level semantic correspondence.

hold. (2) Part-level semantic correspondence. This property is critical for generalization, since the model requires the understanding of semantic relations to make reasonable predictions on novel objects. In addition, good correspondence proves advantages in scenarios with limited data, as the model can be more robust to intra-class recognition, and less susceptible to changes in appearance. We then analyze the features from three representative and powerful visual foundation models, *i.e.*, vision-language contrastive learning CLIP, fully-supervised learning DeiT III [58], and self-supervised learning DINOv2. First, we perform the principal component analysis (PCA) on the extracted patch features of each model to investigate the part awareness. Visualization of PCA components in the top row of Fig. 4 shows that all three models have part-aware features to some extent, yet CLIP cannot well distinguish the background, and features of DeiT III are not discriminative enough for different parts. Next, we choose a different object that has equivalent affordances, *i.e.*, knife and scissors, to assess the semantic correspondence. The bottom row of Fig. 4 shows the feature similarity maps computed as the cosine similarity between one patch representation on the knife blade and an image of scissors. It is obvious that DINOv2 shows finer correspondence between blades of knife and scissors. By contrast, CLIP produces messy correspondences in both foreground and background, and feature correspondences of DeiT III are only discriminative at the object level, but not specific to the affordance part region. From the above analysis, we conclude that DINOv2 is well suited for affordance learning due to its fine-grained part-aware representation and superior part-level semantic correspondence. Quantitative comparisons are shown in Sec. 4.5.

3.3. Motivation and Method

Through a systematic analysis, we identify DINOv2 as a powerful tool for addressing the OOAL problem. However, there are still fundamental issues that hinder performance in this challenging setting. The first is that DINOv2 is a vision-only model, and lacks the ability to identify unseen affordances. One potential solution involves integrating a text encoder like CLIP, but it is recognized that the input text is sensitive to prompts. This is particularly problematic in the case of affordances, which combine both an object and a verb, making manual prompt design a complex task. The second issue is that while features of DINOv2 are part-oriented, the level of granularity varies across layers. Determining the appropriate granularity level is crucial when handling affordances associated with diverse objects. The third issue arises due to the absence of alignment between the DINOv2 vision encoder and CLIP text encoder, as they are trained separately and independently of each other. Building upon these observations, we establish

a vision-language framework based on DINOv2 and CLIP, and propose three modules to resolve each of the three fundamental bottlenecks mentioned above.

In this section, we first describe the overview of our proposed learning framework that builds on the powerful foundation models. Then, we elaborate on the three proposed designs that help in the challenging OOAL problem. Finally, we discuss the framework's capability to identify unseen objects and affordances at inference.

Overview. The proposed learning framework is presented in Fig. 3, which consists of a vision encoder, a text encoder, and a transformer decoder. First, the pretrained vision encoder DINOv2 is used to extract dense patch embeddings $\hat{F}_v \in \mathbb{R}^{L \times C_v}$, where L is the number of tokens or patches. Then, affordance labels are processed by the CLIP text encoder to obtain text embeddings $F_t \in \mathbb{R}^{N \times C}$. To cope with inconsistent dimensions between visual and text embeddings, an embedder $e_v : \mathbb{R}^{C_v} \rightarrow \mathbb{R}^{C}$ with a single MLP layer is employed to transform $\hat{F}_v$ to F_v. In the end, the lightweight transformer decoder takes both visual and text embeddings as input, and outputs the affordance prediction.

Text Prompt Learning. Manually designing prompts for affordances can be a complicated work, especially considering that CLIP has difficulty in recognizing affordance (see Fig. 2). Thus, we adopt the Context Optimization (CoOp) [66] method to introduce automatic text prompt learning. Instead of finetuning the CLIP text encoder, the inclusion of learnable prompts is an effective strategy that can alleviate overfitting and retain the inherent text recognition ability of CLIP. Specifically, p randomly initialized learnable context vectors $\{v_1, v_2, ..., v_p\}$ are inserted in front of the text CLS token, and are shared for all affordance classes.

Multi-Layer Feature Fusion. Different layers of DINOv2 features often exhibit different levels of granularity [1]. Since affordance may correspond to multiple parts of an object, a diverse set of granularities can be beneficial. For this purpose, we aggregate the features of the last j layers. Each layer of features is first processed by a linear projection, and then all features are linearly combined with a weighted summation:

$$\hat{F}_v = \sum_{i=1}^{j} \alpha_i \cdot \phi(F_{n-i+1}), \quad \alpha_1 + \alpha_2 + ... + \alpha_j = 1, \quad (1)$$

where F_n denotes the last layer, α is a learnable parameter that controls the fusion ratio of each layer, and ϕ indicates the linear transformation. This straightforward fusion scheme enables adaptive selection among different granularity levels, allowing the model to handle affordance recognition across diverse scenarios.

CLS-Guided Transformer Decoder. To deal with the lack of alignment between visual and text features, we propose a lightweight transformer decoder that applies a masked cross-attention mechanism to promote the mutual communication between two branches. Since the [CLS] token of a foundation model is used in the computation of objective function, it often carries rich prior information of the whole image, such as salient objects or regions. Consequently, we utilize the [CLS] token to produce a guidance mask that constrains the cross-attention within a foreground region.

The decoder receives three inputs, *i.e.*, text embeddings F_t, visual features F_v, and the [CLS] token L_{cls}. Firstly, linear transformations are performed to yield query, key, and value:

$$Q = \phi_q(F_t), \; K = \phi_k(F_v), \; V = \phi_v(F_v). \quad (2)$$

Here we use text embeddings as query, and visual features as key and value, allowing the model to focus on the update of text embeddings by retrieving relevant visual information that corresponds to the affordance text. Next, the CLS-guided mask is calculated between the [CLS] token and key via matrix multiplication:

$$M_{cls} = \text{sigmoid}(\frac{\phi_c(L_{cls})K^T}{\sqrt{d_k}}), \quad (3)$$

where d_k is a scaling factor that equals the dimension of the keys. The masked cross-attention is then computed as:

$$\hat{F}_t = \text{softmax}(QK^T/\sqrt{d_k}) \cdot M_{cls}V + F_t. \quad (4)$$

After that, the updated text embeddings $F_t^{'}$ are obtained by sending $\hat{F}_t$ through a feed-forward network (FFN) with a residual connection. The decoder comprises t layers of transformers, and the ultimate prediction is generated by performing matrix product between the output of the last transform layer and original visual features F_v, thereby ensuring the maximum retention of part-aware representations from DINOv2. Lastly, binary cross entropy is employed as loss function to optimize parameters of linear layers, embedder, and decoder.

Inference on Unseen Objects and Affordances. During the training process, the decoder learns to establish an alignment between visual features and affordance text embeddings. When encountering a novel object at inference, the aligned affordance text embeddings can locate corresponding object regions, leveraging the part-level semantic correspondence property inherent in DINOv2. Similarly, as the model processes unseen affordance text inputs, the generated text embeddings can also retrieve the aligned visual features, which are based on the semantic similarities to the base affordances seen in the training.

4. Experiments

4.1. Datasets

We choose two typical datasets, AGD20K [35] and UMD part affordance [44], both of which include a large

Task	Training Data seen / unseen split	Method	Seen			Unseen		
			KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
WSAG	23,083 / 15,543 images image-level labels	Hotspots [41]	1.773	0.278	0.615	1.994	0.237	0.577
		Cross-view-AG [35]	1.538	0.334	0.927	1.787	0.285	0.829
		Cross-view-AG+ [34]	1.489	0.342	0.981	1.765	0.279	0.882
		LOCATE [28]	1.226	0.401	1.177	1.405	0.372	1.157
OOAL	50 / 33 images keypoint labels	MaskCLIP [65]	5.752	0.169	0.041	6.052	0.152	0.047
		SAN [61]	1.435	0.357	0.941	1.580	0.351	1.022
		ZegCLIP [67]	1.413	0.387	1.001	1.552	0.361	1.042
		Ours	**0.740**	**0.577**	**1.745**	**1.070**	**0.461**	**1.503**

Table 1. Comparison with state of the art on AGD20K dataset. OOAL setting uses 0.22% / 0.21% of the full training data. WSAG denotes weakly-supervised affordance grounding. The **best** and second-best results are highlighted in bold and underlined, respectively.

Setting	Method	Seen	Unseen	hIoU
Fully Supervised	DeepLabV3+ [10]	70.5	57.5	63.3
	SegFormer [60]	74.6	57.7	65.0
	PSPNet [64]	72.0	**60.8**	66.0
OOAL	PSPNet [64]	56.7	46.6	51.1
	DeepLabV3+ [10]	56.8	48.4	52.3
	SegFormer [60]	64.6	51.4	57.3
	MaskCLIP [65]	4.25	4.24	4.25
	SAN [61]	45.1	32.2	37.5
	ZegCLIP [67]	47.4	36.0	40.9
	Ours	**74.6**	59.7	**66.4**

Table 2. Comparison on UMD dataset. Fully-supervised methods are trained with 14,823 and 20,874 images with pixel-level labels for seen and unseen split, respectively. In contrast, OOAL setting uses 54 and 76 images, 0.36% of the full training data.

number of object categories that help in the evaluation of novel objects. AGD20K is a large-scale affordance grounding dataset with 36 affordances and 50 objects, containing 23,816 images from exocentric and egocentric views. It aims to learn affordance from human-object interaction images, and perform affordance localization on egocentric images. As it is a dataset for weakly-supervised learning, images in the training set only have image-level labels. Therefore we manually annotate 50 randomly selected egocentric images from each object category for training. AGD20K also has two train-test splits for seen and unseen settings, and we follow their splits to evaluate the performance. Note that AGD20K uses sparse annotation, where ground truth consists of keypoints within affordance areas, and then a gaussian kernel is applied over each point to produce dense annotation.

UMD dataset consists of 28,843 RGB-D images with 7 affordances and 105 kitchen, workshop, and gardening tools. It has two train-test splits termed category split and novel split. We use the category split to evaluate base object categories and novel split to evaluate performance on novel object classes. Due to its small number of object categories, we take one example from each base object instance to form the training set. Specific affordance categories and object class splits can be found in the supplementary material.

4.2. Implementation details

Experiments are implemented on two GeForce RTX 3090 GPUs. All visual foundation models use the same base-sized vision transformer (ViT-base). We train the model using SGD optimizer with learning rate 0.01 for 20k iterations. For experiments on AGD20K, images are first resized to 256 × 256 and randomly cropped to 224 × 224 with horizontal flipping. Experiments for UMD dataset are conducted on the opensource toolbox MMSegmentation [13] with the default training setting. The hyperparameters p, j, and t are set to 8, 3, and 2, respectively.

Following previous work, we adopt the commonly used Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) metrics to evaluate the results on AGD20K. For UMD dataset, we use the metric of mean intersection-over-union (mIoU), and also incorporate the harmonic mIoU as a balanced measure that accounts for both seen and unseen settings.

4.3. Comparison to state-of-the-art methods

AGD20K dataset is benchmarked with weakly supervised affordance grounding (WSAG) approaches, which use image-level object and affordance labels to do affordance segmentation. Note that results from WSAG methods are not directly comparable to our setting, as training labels are different. Despite using only image-level labels, the training data required are more than 460 times of ours. The results in Tab. 1 demonstrate that our performance exceeds all WSAG counterparts in an easy and realistic setting. We also benchmark open-vocabulary segmentation methods of MaskCLIP, SAN, and ZegCLIP for further comparison. We find that these CLIP-based methods have a large perfor-

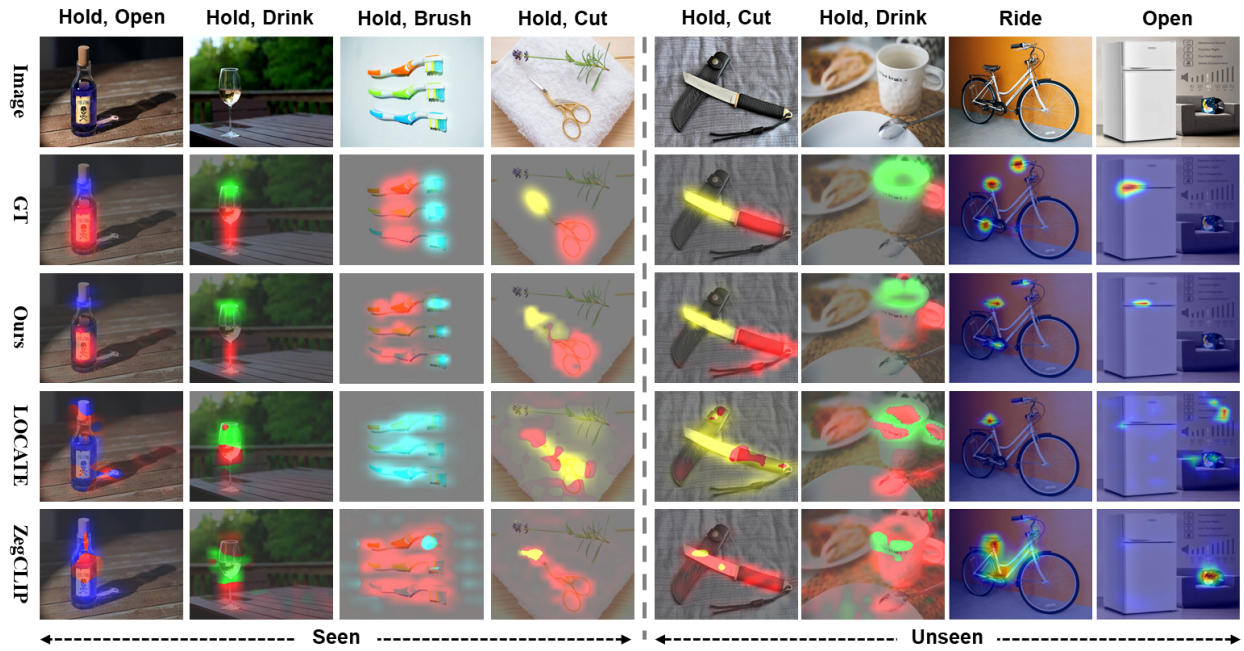

Figure 5. Qualitative comparison with LOCATE and ZegCLIP on AGD20K dataset. When multiple affordance predictions overlap, the one with higher value is displayed. Our predictions distinguish different object parts, while other methods often make overlapping predictions.

Figure 6. Qualitative comparison with SegFormer and ZegCLIP on UMD affordance dataset in OOAL setting. Images have been enlarged and cropped for better visualization.

mance gap with ours, and are also inferior to the state-of-the-art WSAG method LOCATE.

The comprehensive comparison on UMD dataset is displayed in Tab. 2, where we benchmark the results with several representative semantic segmentation methods (PSPNet, DeepLabV3+, SegFormer) and open-vocabulary semantic segmentation methods. For fair comparison, the classical segmentation methods are trained with the full training set, while foundation-model-based methods like ZegCLIP and SAN are evaluated in the OOAL setting. It is clear that our proposed model is quite effective, which can be comparable to fully-supervised methods with only 0.36% of their training data. To explore how fully-supervised methods are affected by the limited data, we further train these models in the OOAL setting. Results in Tab. 2 show that the performance of these models degrades by around 10% in both seen and unseen settings when given only one-shot example. Additionally, under the same OOAL setting, we observe a more apparent gain over other CLIP-based open-vocabulary segmentation methods, showing that CLIP is not suitable for data-limited affordance learning. The poor performance of MaskCLIP from both tables also verifies that vanilla CLIP has limited understanding on affordances.

4.4. Qualitative results

Qualitative comparisons on AGD20K dataset are shown in Fig. 5. We note that WSAG methods like LOCATE often make overlapping predictions for examples with multiple affordances, while our results show a clear separation between different affordance regions. ZegCLIP can make reasonable predictions to some extent, but it mostly focuses

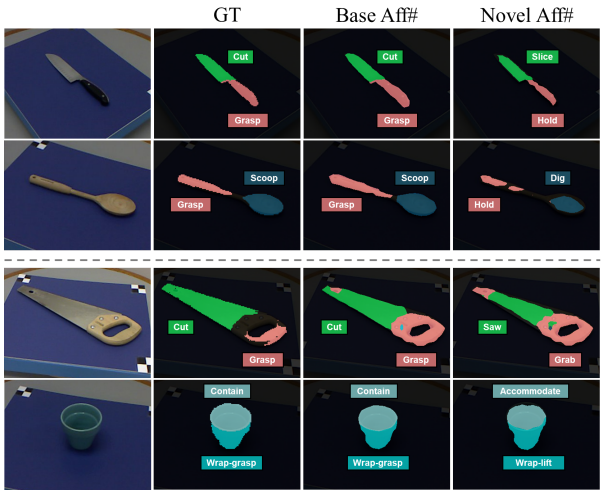

Figure 7. Qualitative examples of novel affordance prediction on UMD dataset. The 1st and 2nd rows display results on base objects, and the 3rd and 4th rows show results for novel objects.

Model	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
CLIP	1.294	0.384	1.107	1.556	0.327	0.966
DeiT III	1.301	0.378	1.140	1.535	0.321	1.049
DINOv2	1.156	0.425	1.297	1.462	0.360	1.105

Table 3. Ablation results of different visual foundation models.

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Baseline	1.156	0.425	1.297	1.462	0.360	1.105
+ TPL	1.060	0.455	1.422	1.338	0.390	1.302
+ MLFF	0.846	0.537	1.622	1.115	0.447	1.440
+ TD	0.749	0.578	1.738	1.131	0.443	1.408
+ CTM	0.740	0.577	1.745	1.070	0.461	1.503

Table 4. Ablation results of proposed modules. TPL: text prompt learning. MLFF: multi-layer feature fusion. TD: transformer decoder. CTM: CLS-guided mask.

on the whole object and the accuracy is far from satisfactory, whereas our results are more part-focused, especially for the unseen objects. For example, the prediction for the unseen object of bicycle show that our model can handle the complex affordance (ride) with multiple separated affordance areas (saddle, handlebar, and pedal). In Fig. 6, we display the results for UMD dataset. We observe that SegFormer and ZegCLIP often fail to recognize affordances of objects whose parts are similar in appearance. Also, they tend to misclassify metallic object parts as cuttable affordance, suggesting that inferring affordances with only appearance features can be misleading. In comparison, our predictions are more accurate due to the utilization of DINOv2's part-level semantic correspondences.

One particular feature of our model is that it can recognize novel affordances not shown during training. To demonstrate this, we replace the original affordance labels with semantically similar words and check if the model can still reason about corresponding affordance areas. As shown in Fig. 7, the model manages to make correct predictions for novel affordances, such as "hold" and "grab" for base affordance "grasp", "saw" for "cut", and "accommodate" for "contain".

4.5. Ablation study

The ablation study is performed on the more challenging AGD20K dataset due to its natural images with diverse backgrounds. Ablations on hyperparameters are left in the supplementary material.

Different Vision Encoders. To complement the qualitative analysis in Sec. 3.2, we conduct quantitative experiments on CLIP, DeiT III, and DINOv2. Specifically, we simply process the visual features with the embedder, and perform matrix multiplication with pre-computed affordance text embeddings to output segmentation maps. As shown in Tab. 3, CLIP and DeiT III exhibit comparable performance, whereas DINOv2 achieves much better results in both seen and unseen settings, which are consistent with the analysis that DINOv2 is more suitable for affordance learning.

Proposed Methods. We use the DINOv2 with a simple embedder as baseline, and gradually integrate our methods to analyze the effect of each proposed design. The results in Tab. 4 reveal that each module can consistently deliver notable improvements. In particular, we notice that the inclusion of a transformer decoder can enhance the performance in the seen setting, but yield inferior results for the unseen setting. With the integration of the CLS-guided mask, results of both settings can be improved, suggesting that restricting the cross-attention space is an effective strategy for unseen object affordance recognition.

5. Conclusion

In this paper, we propose the problem of one-shot open affordance learning that uses one example per base object category as training data, and has the ability to recognize novel objects and affordances. We first present a detailed analysis into different foundation models for the purpose of data-limited affordance learning. Motivated by the analysis, we build a vision-language learning framework with several proposed designs that better utilize the visual features and promote the alignment with text embeddings. Experiment results on two affordance segmentation datasets demonstrate that we achieve comparable performance with less than 1% of the full training data.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For*, 2022. 5

[2] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020. 2

[3] Paola Ardón, Eric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019. 2

[4] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 2

[5] Homanga Bharadhwaj, Abhinav Gupta, and Shubham Tulsiani. Visual affordance prediction for guiding robot exploration. *arXiv preprint arXiv:2305.17783*, 2023. 2

[6] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6372–6378. IEEE, 2022. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[9] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023. 2

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6

[11] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*, 2023. 3

[12] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 1, 2

[13] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[14] 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. 1

[15] 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. 1, 2

[16] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, 2021. 1

[17] Thanh Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. *ICRA*, 2018. 1, 2

[18] Kuan Fang, Te Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2Vec: Reasoning Object Affordances from Online Videos. *CVPR*, 2018. 1, 2

[19] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. End-to-end affordance learning for robotic manipulation. *arXiv preprint arXiv:2209.12941*, 2022. 1

[20] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979. 1

[21] Denis Hadjivelichkov, Sicelukwanda Zwane, Marc Deisenroth, Lourdes Agapito, and Dimitrios Kanoulas. One-Shot Transfer of Affordance Regions? AffCorrs! *CoRL*, 2022. 2

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

[23] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 1

[24] 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. *arXiv preprint arXiv:2307.05973*, 2023. 2

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[26] Mia Kokic, Johannes A Stork, Joshua A Haustein, and Danica Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE, 2017. 1, 2

[27] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 2

[28] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023. 1, 2, 6

[29] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-

vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 3

[30] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 1, 2

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[32] Timo Luddecke and Florentin Worgotter. Learning to segment affordances. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 769–776, 2017. 1

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

[34] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *arXiv preprint arXiv:2208.13196*, 2022. 6

[35] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. *CVPR*, 2022. 1, 2, 5, 6

[36] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11576–11582. IEEE, 2023. 2

[37] Reihaneh Mirjalili, Michael Krawez, Simone Silenzi, Yannik Blei, and Wolfram Burgard. Lan-grasp: Using large language models for semantic object grasping. *arXiv preprint arXiv:2310.05239*, 2023. 2

[38] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and Jose Santos-Victor. Affordances, development and imitation. In *2007 IEEE 6th International Conference on Development and Learning*, pages 270–275. IEEE, 2007. 1

[39] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Bayesian deep learning for affordance segmentation in images. *arXiv preprint arXiv:2303.00871*, 2023. 1

[40] Austin Myers, Angjoo Kanazawa, Cornelia Fermuller, and Yiannis Aloimonos. Affordance of Object Parts from Geometric Features. *Int. Conf. Robot. Autom.*, pages 5–6, 2015. 2

[41] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1, 2, 6

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

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

[44] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 1, 5

[45] Toan Ngyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. *arXiv preprint arXiv:2303.02401*, 2023. 2

[46] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 2

[47] 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. 3

[48] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. *arXiv preprint arXiv:2401.06341*, 2024. 2

[49] 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. 2, 3

[50] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. *arXiv preprint arXiv:2309.07970*, 2023. 2

[51] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 186–201. Springer, 2016. 1

[52] Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *ICCVW*, 2017. 1, 2

[53] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. *CVPR*, 2017. 1, 2

[54] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 3

[55] Yaoxian Song, Penglei Sun, Yi Ren, Yu Zheng, and Yue Zhang. Learning 6-dof fine-grained grasp detection based on part affordance grounding. *arXiv preprint arXiv:2301.11564*, 2023. 2

[56] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. *arXiv preprint arXiv:2305.11173*, 2023. 3

[57] Chao Tang, Jingwen Yu, Weinan Chen, and Hong Zhang. Relationship oriented affordance learning through manipulation graph construction. *arXiv preprint arXiv:2110.14137*, 2021. 2

[58] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, pages 516–533. Springer, 2022. 4

[59] Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. *arXiv preprint arXiv:2310.05107*, 2023. 3

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6

[61] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 6

[62] Xintong Yang, Ze Ji, Jing Wu, and Yu-Kun Lai. Recent advances of deep robotic affordance learning: a reinforcement learning perspective. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. 2

[63] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. *arXiv preprint arXiv:2303.10437*, 2023. 2

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 6

[65] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3, 6

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5

[67] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 6