# MaskPrompt: Open-Vocabulary Affordance Segmentation with Object Shape Mask Prompts

**Dongpan Chen, Dehui Kong*, Jinghua Li, Baocai Yin**

School of Information Science and Technology, Beijing University of Technology, Beijing, China
cdp@emails.bjut.edu.cn, {kdh, lijinghua, ybc}@bjut.edu.cn

## Abstract

Affordance refers to the interactable functional properties of an object, and affordance segmentation aims to pixel-level segment the object functional parts in a given image, which is crucial for various interactive vision tasks. Existing methods address the affordance segmentation problem by utilizing only image features, they can hardly solve the problems of interference between adjacent object pixels in complex scenes, and inability to generalize to the open-world. To tackle these problems, we propose a novel open-vocabulary affordance segmentation task and a benchmark dataset, and propose an approach with object shape mask prompts. The mask is used as prior for different granularity visual feature enhancement and fine-grained text prompt embedding. Specifically, we first propose a mask prompt generation module, which generates refined object shape masks, as well as text prompts for mask-focused regions. Based on the masks, we propose a mask prompt feature enhancement module. It uses masks to encode instance features, and then aggregates them with global features to enhance the visual feature representation. The enhanced visual features are combined with text prompts of different granularity to generate class-agnostic affordance mask proposals. We finally classify these proposals in a proposed affordance prediction module. Quantitative and qualitative evaluations compared with state-of-the-art methods demonstrate that the proposed method achieves superior performance on the proposed benchmark dataset and other open-vocabulary part segmentation datasets.

## Introduction

Affordance segmentation aims to segment the functional parts of an object in a given image. But in real life, many vision tasks need to segment not only the known object affordances in the training datasets, but also the unknown affordances. To our knowledge, no studies have been conducted to address these issues. Thus we propose an novel open-vocabulary affordance segmentation (OVAS) task. OVAS not only segments the affordances of objects in closed datasets, but also extends to the open-world by segmenting affordances that did not occur in the training process. OVAS plays an important role in many computer vision fields, such as scene understanding (Balazevic et al. 2024), human-object interaction (Wu et al. 2024), and especially in robotics (Xu et al. 2024a; Ragusa et al. 2023), where it segments interactable regions for robots to perform interaction tasks.


Figure 1: Examples of open-vocabulary affordance segmentation results, and open-vocabulary affordance categories.

However, open-vocabulary affordance segmentation is a very challenging task. As shown in Fig. 1, on the one hand, affordance segmentation is a part-level image segmentation task, affordance is the fine-grained version of an object. Different from traditional object-level semantic segmentation, object parts have complex structures and usually have more complex boundaries and appearance variations than the object as a whole. In common chaotic scenes, the interference of background and adjacent objects' pixels significantly affect the quality of visual feature encoding, leading to unsatisfactory segmentation result. On the other hand, open-vocabulary affordance segmentation introduces an open granularity challenge. OVAS requires segmenting the affordances did not exist in the training datasets, this is difficult for data-driven deep learning segmentation models. It requires a lot of additional information supplementation, as well as effective multi-modal fusion.

Some related work has been proposed attempt to solve the above problems. One kind of methods such as (Minh et al. 2020; Gu, Su, and Yuan 2021; Lu et al. 2022), which only encode the global features of the image in an end-to-end manner. However, they are not effective in suppressing interference from complex background. Another kind of methods (Nguyen et al. 2017; Do, Nguyen, and Reid 2018) training an object detector to obtain object region features for affordance segmentation. These methods can mitigate the back-

*Corresponding author.

ground interference to some extent, but can not reduce the interference of the pixels of parts in adjacent objects well. To solve this problem, Chen *et al.* (Chen et al. 2024) propose an affordance segmentation network with object mask guided feature encoders, the mask acts as an attention mechanism to make the network focus on the features of the object region. This model improves the segmentation performance, but it relies on the accuracy of mask generator. For open-vocabulary affordance segmentation, there is currently no direct research. However, open-vocabulary affordance segmentation is similar to open-vocabulary part segmentation at the technical level, it can be executed using open-vocabulary part segmentation methods. These methods (Pan et al. 2023; Sun et al. 2023a; Wei et al. 2024) combine the mask generation model (e.g., MaskFormer (Cheng, Schwing, and Kirillov 2021)) with large-scale vision-language models (VLMs) like CLIP (Radford et al. 2021) in a two-stage manner. In the first stage, the mask generation models generate class-agnostic mask proposals. In the second stage, the CLIP model classifies these proposed part masks.

To address the problems faced by open-vocabulary affordance segmentation task, i.e. the interference of complex background and pixels of adjacent objects, as well as the problem of generalization to the open-world, inspired by work (Chen et al. 2024) and open-vocabulary part segmentation, we propose a benchmark dataset and a novel open-vocabulary affordance segmentation method with object shape mask prompts, namely MaskPrompt. To solve the problem that traditional affordance segmentation methods do not model image features well, we propose to generate refined object shape masks, which are used as prompts to generate masked images containing only single object instance. This can exclude the background interference and compensate the object features well, thus enhancing the overall visual features. As for the open-world affordance segmentation problem, we use masks to generate fine-grained captions as text prompts, which are used to generate high-quality mask proposals, and further classify the masks accurately.

Specifically, the proposed MaskPrompt consists of three main modules, that is, mask prompt generation module (MPGM), mask prompt feature enhancement module (MPFEM), and affordance prediction module (APM). In MPGM, we first generate the object shape mask. Given an input image, we use a pre-trained object detector (such as DETR (Carion et al. 2020)) to detect the object categories. Then the object categories and the image are fed into an image segmentation model, i.e., segment anything model (SAM) (Kirillov et al. 2023), to get the refined target object shape masks. Further, to obtain fine-grained text prompts, we input the mask and original image into Alpha-CLIP (Sun et al. 2023b) to output caption descriptions specific to the mask region. The Alpha-CLIP is a variant of the CLIP that allows you to get information about wherever you want to focus on. Since we focus primarily on the objects, we can get text captions specific to the target objects. We use these captions and object categories, as well as the open affordance words, as joint text prompts and embed them into the feature space using a text encoder. In MPFEM, we use mask to remove the background and get the masked image containing only a single object, which is used to extract the object instance features. Different from the traditional affordance segmentation methods based on object detection, the detector extracts the object features form rectangular boxes, which also contains some background information in addition to the objects. In contrast, the masked images contain information of only object shape region, which can provide more direct and valuable features for the model. In APM, the encoded instance features are concatenated with the global image features, and then fed together with the text prompting features into a proposed multi-modal pixel decoder, which outputs the refined class-agnostic affordance mask proposals. These proposals are multiplied with the text embedding to obtain the affordance classes of the masks. We get the final open-vocabulary affordance segment map by combining the affordance classes and masks.

Overall, our contributions are summarized as follows:

- We propose a novel open-vocabulary affordance segmentation task, and a corresponding annotated benchmark dataset. To the best of our knowledge, this is the first task proposed to address the affordance segmentation problem in the open-world of real scenes.
- We propose MaskPrompt, a novel open-vocabulary affordance segmentation method based on object shape mask prompts. MaskPrompt utilizes generated fine-grained masks to enhance instance features and embed prompts of different granularity, continuously suppressing interference and improving generalization ability.
- We conduct extensive experiments on the benchmark and other object part segmentation datasets, which demonstrates the effectiveness of our proposed method.

## Related Work

### Affordance Segmentation

Traditional affordance segmentation methods are based on the handcrafted low-level features of the objects, but they lack the powerful learning ability of deep learning methods and has poor performance. Some deep learning based methods mine visual features at different levels (Roy and Todorovic 2016; Do, Nguyen, and Reid 2018; Minh et al. 2020; Bastanfard, Amirkhani, and Mohammadi 2022; Mur-Labadia, Martinez-Cantin, and Guerrero 2023), or use attention mechanism (Zhao, Cao, and Kang 2020; Gu, Su, and Yuan 2021; Wang and Tian 2023) for affordance segmentation. In addition, weakly supervised methods are popular in affordance segmentation because they reduce the dependency on annotated datasets (Sawatzky and Gall 2017; Chu, Xu, and Vela 2019). Affordance segmentation in 3D data also has been explored. Deng *et al.* (Deng et al. 2021) propose a 3D point cloud affordance segmentation dataset based on PartNet (Mo et al. 2019) and ShapeNet (Chang et al. 2015). Xu *et al.* (Xu et al. 2022) propose part-level affordance segmentation from 3D Objects. Nguyen *et al.* (Nguyen et al. 2023) propose utilizing affordance labels as prompts for 3D affordance segmentation. Although these methods have designed complex network structures to improve the segmentation performance, they only focus on the global features and

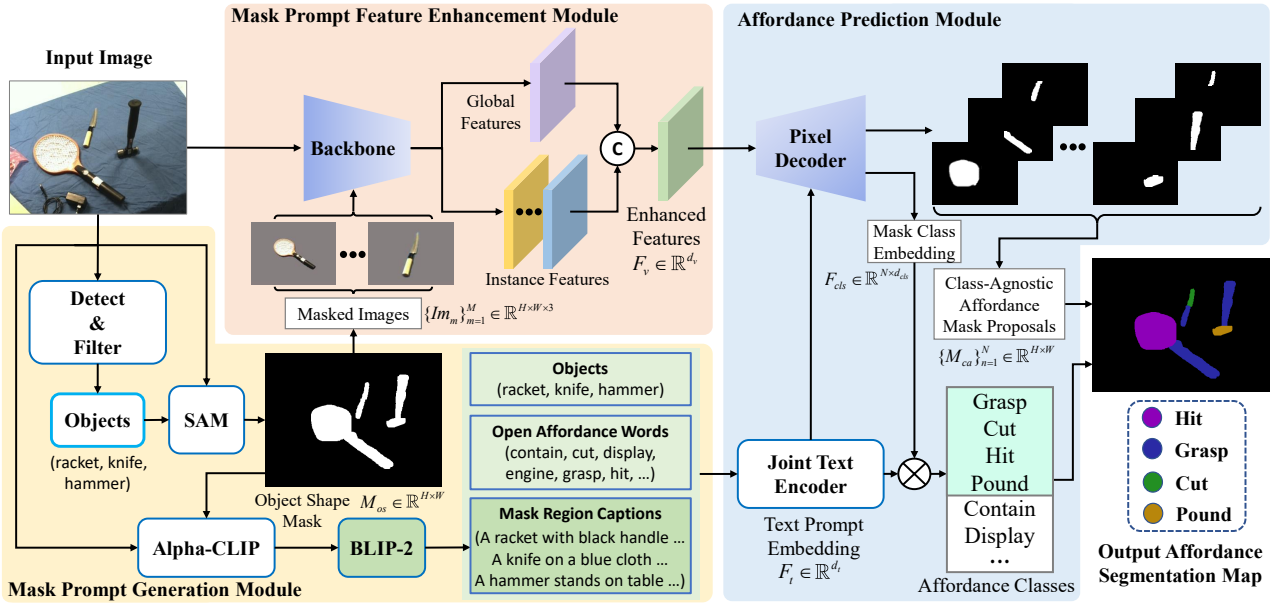


Figure 2: The framework of MaskPrompt. Given an input image, the mask prompt generation module generates refined object shape masks and fine-grained mask region captions. The mask prompt feature enhancement module uses the masks to encode object instance features to enhance the feature representation. The enhanced features and joint text prompt embeddings are cooperatively fed into a proposed pixel decoder to generate the refined class-agnostic affordance mask proposals. The affordance prediction module gets the final predicted affordance segmentation map by multiplying the mask proposals and text embeddings.

do not make fully exploit the individual objects to enhance the feature representation, which makes the interference of neighbour objects a difficult problem to solve.

## Open-Vocabulary Image Segmentation

Open-vocabulary image segmentation aims to segment the target categories that did not appear in the training process, to improve its generalisation to the open-world. Existing open-vocabulary segmentation models mainly map visual features to semantic space (Yue et al. 2024; Xu et al. 2024b; He 2024), or leverages large pretrained vision-language models (such as CLIP) to distill or transfer the visual-semantic knowledge (Ding et al. 2022; Li et al. 2021; Liu et al. 2023; Wu et al. 2023; Han et al. 2023). Bucher *et al.* (Bucher et al. 2019) generates the pixel-level features of unseen classes in the semantic embedding space and adopts the generated features to supervise a visual segmentation model. Qin *et al.* (Qin et al. 2023) propose a generic framework to accomplish unified, universal and open-vocabulary image segmentation. They propose adaptive prompt learning that facilitates the unified model to capture task-aware and category-sensitive concepts, improving model robustness in multi-task and varied scenarios. Zhang *et al.* (Zhang et al. 2023) propose a simple framework for open-vocabulary segmentation and detection. They identify the discrepancies in two tasks and propose separate techniques including shared semantic space, decoupled decoding, and conditioned mask assistance to mitigate the issues. Han *et al.* (Han et al. 2023) propose global knowledge calibration to preserve generalizable representations when training solely on known classes. Inspired by these successful works, we introduce text prompts into our research.

# Method

## Overview

Open-vocabulary affordance segmentation aims to train a model $f_\theta(\cdot)$ on a training dataset $D_{train}$, and then test on a test dataset $D_{test}$ for both training base categories $C_{base}$ and novel categories $C_{novel}$ of affordance. $f_\theta(\cdot)$ segment the fine-grained functional parts of objects in a given image. The novel affordance categories are not visible in the training dataset, but are visible in the test dataset. Formally, for an input image $I \in \mathbb{R}^{h\times w\times 3}$, we aim to identify and segment the functional parts of the target entities, i.e., $\{(y_n, m_n\}_{n=1}^{N}$, where $m \in \{0,1\}^{H\times W}$ represents the binary mask of a part and $y$ denotes its corresponding affordance class, assuming the entity contains $N$ parts, and some of parts may have the same affordance class label.

To achieve the goal, we propose a novel open-vocabulary affordance segmentation method via object shape mask prompts, namely MaskPrompt, which is illustrated in Fig. 2. MaskPrompt mainly consists three modules, i.e., mask prompt generation module (MPGM), mask prompt feature enhancement module (MPFEM), and affordance prediction module (APM). The MPGM generates refined object shape masks, as well as text prompts for mask-focused regions. The MPFEM uses masks to encode instance features, and then aggregates them with global features to enhance the visual feature representation. The enhanced visual features are combined with text prompts of different granularity to

generate class-agnostic affordance mask proposals. Finally APM classifies these masks and outputs the final affordance segmentation map.

## Mask Prompt Generation

In this module, we mainly utilize object shape mask to obtain fine-grained text prompts, as well as remove the background of the original image and generate masked images containing object instances. Therefore, we first need to generate refined object shape mask. In order to reduce model complexity and computational overhead, we utilize an off-the-shelf pre-trained image segmentation model, segment anything model (SAM) (Kirillov et al. 2023). SAM produces high quality object masks from input prompts such as text or boxes, and it can be used to generate masks for all objects in an image. It has been trained on a dataset of 11 million images and 1.1 billion masks, and has strong zero-shot performance on a variety of segmentation tasks. Therefore, it is necessary to detect the target object categories, and SAM can accurately segment the object shape mask based on the object category prompts. We use DETR (Carion et al. 2020), an end-to-end transformer-based object detection model. Different from traditional candidate box based methods, DETR does not require pre generation of candidate boxes. It directly generates a fixed number of object queries from the image and infers the entire image through self attention mechanism, thereby avoiding the process of candidate box generation and non maximum suppression, as well as simplifying the process. Specifically, the process of object detection is described as follows:

$$S_{obj} = \mathrm{DETR}(I; Q), \tag{1}$$

where $S_{obj}$ represents the detected object scores, $I$ is the input image, and $Q$ is the learnable queries initialized in DETR. Since the objects studied in this paper are all included in the pre-trained DERT weights, there is no need to train them separately. We choose to keep the results with predicted scores greater than a threshold $T$, and finally remove the objects that are not in the set of target objects $O_t$ we are studying. The specific process of obtaining the final object sets $O_s$ are described as follows:

$$O_s = \{S_{obj} > T | Class(obj) \in O_t\}, \tag{2}$$

where $Class(obj)$ represents the word of detected object.

The process of SAM segmentation is described as follows:

$$M_{os} = \mathrm{SAM}(I; Class(obj) \in O_s), \tag{3}$$

where $M_{os}$ represents the binary object shape mask, $I$ is the input image, and $obj$ are the detected target objects.

In addition, considering the success of image-text contrastive learning, we propose using text prompts to guide the affordance mask prediction. The existing visual task related methods based on text prompts all utilize large-language models such as BLIP-2 (Li et al. 2023) to generate text descriptions. However, BLIP-2 generates all text information containing the entire image, but for specific tasks, only some of the information is valuable, while others become noise. To this end, we propose using an existing novel model, Alpha-CLIP (Sun et al. 2023b), it takes the object mask as input and outputs a text description of the mask region. The process of Alpha-CLIP text prompt generation as follows:

$$w_{mask} = \text{BLIP-2}\{\text{Alpha-CLIP}(\mathrm{I}; \mathrm{M_{os}})\}, \tag{4}$$

where $w_{mask} = \{sen_1, sen_2, \cdots, sen_n\}$ ($sen$ is the generated sentences, $n$ is the number of masks) represents the text prompts about the object shape mask ($M_{os}$) region.

## Mask Prompt Feature Enhancement

Conventional visual feature encoders only encode global information about the images, but for the tasks of affordance segmentation, these features are insufficient to reduce the interference of background and pixels of adjacent target entities. Therefore, we propose adding additional instance features to enhance feature representation.

Specifically, we mask the original input image $I \in \mathbb{R}^{H \times W \times 3}$ based on the object shape mask ($M_{os} \in \mathbb{R}^{H \times W}$) obtained by MPGM. The masked images ($\{Im_m\}_{m=1}^{M} \in \mathbb{R}^{H \times W \times 3}$, where $m$ denotes the masked image number) remove background information and contain only one instance, and pixels in other areas are filled with the average of image pixels. These images and the original image are fed into the backbone to obtain masked features $\{Fm_m\}_{m=1}^{M}$ and original features $F_o$, respectively. We then concatenate these features and reduce the number of channels through a convolution layer to obtain the enhanced visual features $F_v$. Specifically, $F_v$ can be described as follows:

$$F_v = C(Cat[F_o; \{Fm_m\}_{m=1}^{M}]) \tag{5}$$

where $Cat[\cdot]$ denotes channel concatenation, and $C(\cdot)$ denotes channel convolution with $1 \times 1$ kernel.

## Affordance Prediction

We adopt a two-stage segmentation pipeline to solve the open-world affordance segmentation problem. The first stage generates class-agnostic affordance mask proposals, and then the second stage classifies these proposals. We first design a transformer based visual-text cross-modal pixel decoder to generate the proposals, which is illustrated in Fig. 3. It takes two inputs, i.e., visual features and text embedding, and outputs a set of affordance mask proposals.

We first transform the text prompts from mask prompt generation module into the embedding space. To add more text descriptions, we introduce words of objects ($w_{obj} = \{obj_1, obj_2, \cdots, obj_\omega\}$, $\omega$ is the number of detected object categories), and all affordances ($w_{aff} = \{aff_1, aff_2, \cdots, aff_C\}$). The former provides features of instances, and the latter provides all affordance classes, as well as categories that do not exist in the training set, which enables open-world affordance segmentation. We use CLIP's text-encoder to convert these syndicated text captions into the embedded space. The specific description is as follows:

$$text_{join} = \mathrm{CLIP.tokenize}(w_{obj}; w_{aff}; w_{mask}), \tag{6}$$

$$F_t = \mathrm{CLIP.encode\_text}(text_{join}), \tag{7}$$

where $text_{join}$ denotes the joint text tokens by "tokenize" function of CLIP, and $F_t$ denotes the joint text prompt embedding by "encode_text" function of CLIP.

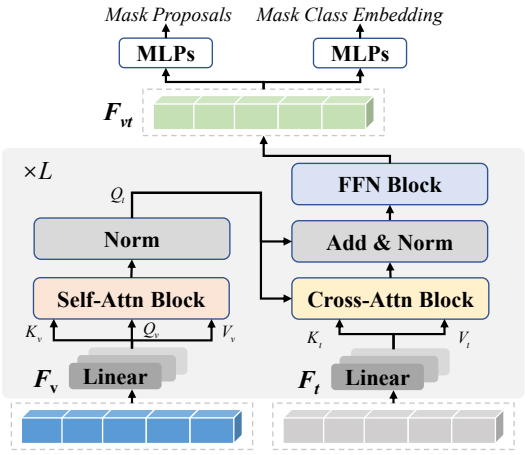


Figure 3: The framework of proposed pixel decoder.

Afterwards, we input the visual features $F_v \in \mathbb{R}^{d_v}$ and text embedding $F_t \in \mathbb{R}^{d_t}$ into the proposed pixel decoder. The visual features $F_v$ are subjected to self-attention block, then cross-attention block with the text features $F_t$, afterwards the fused visual-text attention features are output through the feed-forward network (FFN) block, which consists of two fully-connected layers with a nonlinear mapping GELU inserted in-between. The final refined features $F_{vt}$ are output after $L$ iterations. The detailed operations are described as follows:

$$Q_v, K_v, V_v = Linear(F_v), \tag{8}$$

$$K_t, V_t = Linear(F_t), \tag{9}$$

$$Q_t = Norm(SA(Q_v; K_v; V_v)), \tag{10}$$

$$F_{vt} = \{FFN(Q_t + Norm(CA(Q_t; K_t; V_t)))\}_{l=1}^{L}, \tag{11}$$

where $Q_v = F_v W_q^v$, $K_v = F_v W_k^v$, $V_v = F_v W_v^v$, $K_t = F_t W_k^t$, $V_t = F_t W_v^t$ in which $W_q^v$, $W_k^v$, $W_v^v$, $W_k^t$, and $W_v^t$ are all learnable linear transformations with the same size $d$. $Norm$ represents L2 normalization. $SA$ and $CA$ indicate self-attention and cross-attention respectively, and they follow a uniform formula, i.e., $softmax(QK^T/\sqrt{d})V$.

Finally, $F_{vt} \in \mathbb{R}^{d_{vt}}$ are fed into MLP layers to generate the class-agnostic affordance mask proposals $\{M_{ca}\}_{n=1}^{N} \in \mathbb{R}^{H \times W}$ and mask class embedding $F_{cls} \in \mathbb{R}^{N \times d_{cls}}$, respectively. Then we use the dot-product between mask class embedding and text embedding to obtain the classification score $s_{cls} \in \mathbb{R}^{N \times C}$ for a open set of affordance classes $C$. The classification score $s_{cls}$ is calculated as follows:

$$s_{cls}^{i,j} = \sigma(\cos(F_{cls}^i; F_t^j)/\varphi), \tag{12}$$

where $i \in [1, N]$ and $j \in [1, C]$ are indexes of object affordance and text embedding respectively, $\sigma(\cdot)$ is the sigmoid function, $cos(\cdot;\cdot)$ denotes the cosine similarity, and $\varphi$ is the temperature hyper-parameter.

## Loss Function

The goal of open-vocabulary affordance segmentation is to generate refined affordance mask proposals and match them with their corresponding ground-truth labels. Formally, the model is optimized by minimizing the loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{cls}(\hat{s}_{cls}; s_{cls}) + \lambda \mathcal{L}_{mask}(\hat{m}; m), \tag{13}$$

where $\mathcal{L}_{cls}$ denotes the binary cross-entropy loss, the mask loss $\mathcal{L}_{mask}$ is the sum of dice loss and binary focal loss, $\hat{s}$ represents ground truth affordance class label, and $\hat{m}$ denotes the ground truth affordance mask label.

# Experiments

## Datasets and Evaluation Metrics

We combine existing affordance segmentation dataset **IIT-AFF** (Nguyen et al. 2017) and part segmentation dataset **Pascal-Part-108** (Michieli et al. 2020), and re-annotate labels according to the affordances of target entities (including objects, humans and animals) to construct an open-vocabulary affordance segmentation dataset, namely **OVAS-25**. OVAS-25 has 28 entity classes and 25 affordance classes (as shown in Fig. 1), totalling 18938 images, of which 11363 are used for training and 7575 for testing. IIT-AFF includes 10 object categories and 9 affordance categories, and consists of 8835 real-world images, of which 6496 images are from the ImageNet (Russakovsky et al. 2015) dataset, and the other 2339 images are from the video frame data in a complex scene collected by a robot camera. To enrich the data samples and better fit the open-world, we collect images from the commonly used part segmentation dataset Pascal-Part-108 and change its annotations. Pascal-Part-108 consists of 20 object categories, 108 object part categories, and a total of 10103 images. We also evaluate the proposed model on another affordance segmentation dataset **UMD** (Myers et al. 2015) and other part segmentation datasets, i.e., **Pascal-Part-58** (Chen et al. 2014), **Pascal-Part-116** (Wei et al. 2024), **Pascal-Part-201** (Singh et al. 2022), and **ADE20K-Part-234** (Wei et al. 2024).

Following the previous work (Singh et al. 2022; Chen et al. 2024), we use the mean of intersection over union (**mIoU**), mean average (**mAvg**), and **F1-Score** to measure the performance of our models.

## Implementation Details

We adopt a pre-trained DETR as the object detector with the threshold $T$ of 0.7, and the parameters of DETR, SAM and Alpha-CLIP are frozen. We train the whole model for 120K iterations with a learning rate of $10^{-4}$ decreased by 10 times at 60K and 100K iterations. We optimize the network by AdamW with the weight decay $10^{-4}$ and batch size 32. The layers of pixel decoder $L$ is 6. In each layer, the embedding dimension is 768, the head number of the multi-head attention is 12, $d$ is 512, and the hidden dimension of the feed-forward network is 3072. For the dimensions of text and vision features, $d_t$, $d_v$, $d_{vt}$, and $d_{cls}$ are all 512. All experiments are conducted on a NVIDIA A800 80GB GPU.

## Comparison Results

**Quantitative Comparison.** Since we address the open-vocabulary affordance segmentation problem, which is essentially the same as open-vocabulary part segmentation

| Method | Backbone | Open/Closed-vocabulary | OVAS-25 | | | IIT-AFF | UMD |
|---|---|---|---|---|---|---|---|
| | | | mIoU | mAvg | F1-Score | F1-Score | F1-Score |
| FLOAT (Singh et al. 2022) | ResNet-101 | Open | 64.65 | 65.07 | 74.17 | - | - |
| OPS (Pan et al. 2023) | ResNet-50 | Open | 64.77 | 65.23 | 74.21 | - | - |
| VLPart (Sun et al. 2023a) | ResNet-50 | Open | 65.84 | 66.14 | 75.80 | - | - |
| OV-PARTS (Wei et al. 2024) | ResNet-101 | Open | 65.99 | 66.28 | 76.06 | - | - |
| GSE (Zhang et al. 2022) | ResNet-101 | Closed | - | - | - | 82.33 | 85.50 |
| RelaNet (Zhao, Cao, and Kang 2020) | ResNet-50 | Closed | - | - | - | 78.92 | 82.94 |
| BPN (Yin and Zhang 2022) | ResNet-50 | Closed | - | - | - | 79.64 | 86.21 |
| ADOSMNet (Chen et al. 2024) | ResNet-101 | Closed | - | - | - | 85.85 | 91.81 |
| Ours | ResNet-50 | Open/Closed | 69.52 | 69.68 | 79.36 | 87.24 | 92.41 |
| Ours | ResNet-101 | Open/Closed | **71.26** | **71.67** | **81.58** | **89.46** | **93.83** |

Table 1: Comparison results of open/closed-vocabulary affordance segmentation methods on OVAS-25, IIT-AFF, and UMD datasets. The best performances are in bold.

| Method | Backbone | Pascal-Part-58 | | Pascal-Part-108 | | Pascal-Part-201 | |
|---|---|---|---|---|---|---|---|
| | | mIoU | mAvg | mIoU | mAvg | mIoU | mAvg |
| BSANet (Zhao et al. 2019) | ResNet-101 | 58.2 | 58.9 | 45.9 | 48.4 | 28.5 | 38.7 |
| GMNet (Michieli et al. 2020) | ResNet-101 | 59.0 | 61.8 | 45.8 | 50.5 | 22.5 | 33.2 |
| CO-Rank (Tan et al. 2021) | ResNet-101 | 60.7 | 60.6 | - | - | - | - |
| FLOAT (Singh et al. 2022) | ResNet-101 | 61.0 | 64.2 | 48.0 | 53.0 | 37.1 | 46.9 |
| Ours | ResNet-101 | **63.8** | **66.4** | **49.5** | **55.7** | **40.6** | **48.7** |

Table 2: Comparison results with open-vocabulary part segmentation methods on Pascal-Part-58, Pascal-Part-108, and Pascal-Part-201 datasets. The best performances are in bold.

tasks. Therefore, we first select the existing state-of-the-art open-vocabulary part segmentation methods, train them on the proposed OVAS-25 dataset, and report the test results as shown in Table 1. Compared to the state-of-the-art open-vocabulary part segmentation method OV-PARTS (Wei et al. 2024), which utilizes learnable text prompts as well as fine-tuned CLIP to improve performance, our method improves by 5.27% mIoU, demonstrating the effectiveness of our proposed mask prompts. In addition, we conduct experiments on the traditional closed-vocabulary affordance segmentation datasets IIT-AFF and UMD, and our method can be directly migrated, which are illustrated in Table 1. Compared with the best methods (Zhang et al. 2022; Zhao, Cao, and Kang 2020; Yin and Zhang 2022; Chen et al. 2024), our model remains competitive.

Moreover, to illustrate the scalability and robustness of our method, we train our model on some other open-vocabulary part segmentation datasets (e.g., Pascal-Part-58, Pascal-Part-108, Pascal-Part-116, Pascal-Part-201, and ADE20K-Part-234) and compare the results with the existing methods as shown in Table 2 and Table 4. In Table 2, compared to the best open-vocabulary part segmentation method FLOAT (Singh et al. 2022), our mIoU score on Pascal-Part-58 and Pascal-Part-201 datasets even surpasses FLOAT. And our proposed method is optimal compared to the others. In Table 4, according to the OV-PARTS (Wei et al. 2024) settings, we report the zero-shot performance of the part segmentation methods on Pascal-Part-116 and ADE20K-Part-234 datasets. Besides, also using ViT as backbone, our method is far better than CLIPseg (Lüddecke and Ecker 2022) and CATSeg (Cho et al. 2023). And using the ResNet-101c backbone, our method is mostly optimal in the unseen setting compared to ZSseg+ (Wei et al. 2024). These experiments indicate that our proposed model has strong robustness in the open-world.

| Configuration | | | Metrics | | |
|---|---|---|---|---|---|
| MPFEM | MPGM | APM | mIoU | mAvg | F1-Score |
| - | - | - | 60.51 | 61.20 | 71.05 |
| ✓ | - | - | 67.41 | 67.87 | 76.98 |
| ✓ | ✓ | - | 69.65 | 70.31 | 79.62 |
| ✓ | ✓ | ✓ | **71.26** | **71.67** | **81.58** |

Table 3: Ablation studies on different proposed modules. The best performances are in bold.

**Qualitative Comparison.** The qualitative comparison is illustrated in Fig. 4. Compared with other open-vocabulary part segmentation methods, our proposed method demonstrates the best visualization results. As shown in the second and third rows, our proposed method can achieve optimal performance in solving complex background interference and adjacent object pixel interference problems. In addition, our method also segments object parts that do not appear in the ground-truth, such as the small bottle as shown in the last row. These results demonstrate the effectiveness of our proposed method, especially for tiny object parts, such as "contain" of the cap as shown in the third row.

## Ablation Study

We conduct ablation studies of different modules on the proposed dataset, which are shown in Table 3. Our basic model only utilizes the original features of the image, while the pixel decoder adopts a structure similar to MaskFormer (Cheng, Schwing, and Kirillov 2021). As shown in the second row of Table 3, we add the mask prompt feature enhancement

| Method | Backbone | Pascal-Part-116 | | | | ADE20K-Part-234 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Oracle-Obj | | Pred-Obj | | Oracle-Obj | | Pred-Obj | |
| | | Seen | Unseen | Seen | Unseen | Seen | Unseen | Seen | Unseen |
| ZSseg+ (Wei et al. 2024) | ResNet-50 | 55.33 | 19.17 | 54.23 | 17.10 | 43.19 | 27.84 | 21.30 | 5.60 |
| ZSseg+ (Wei et al. 2024) | ResNet-101c | 57.88 | 21.93 | 56.87 | 20.29 | 43.41 | 25.70 | 21.42 | 3.33 |
| CATSeg (Cho et al. 2023) | ViT-B/16 | 43.97 | 26.11 | 41.65 | 26.08 | 31.40 | 25.77 | 20.23 | 8.27 |
| CLIPSeg (Lüddecke and Ecker 2022) | ViT-B/16 | 48.68 | 27.37 | 44.57 | 27.79 | 38.96 | 29.65 | 24.80 | 6.24 |
| Ours | ResNet-101c | 58.37 | 28.36 | 57.10 | 28.62 | 44.38 | 31.21 | 24.95 | 8.77 |
| Ours | ViT-B/16 | **61.24** | **30.81** | **59.74** | **29.87** | **46.65** | **33.20** | **25.54** | **9.81** |

Table 4: Zero-shot performance of the part segmentation methods on Pascal-Part-116 and ADE20K-Part-234. The evaluation metric is mIoU and the best performances are in bold.

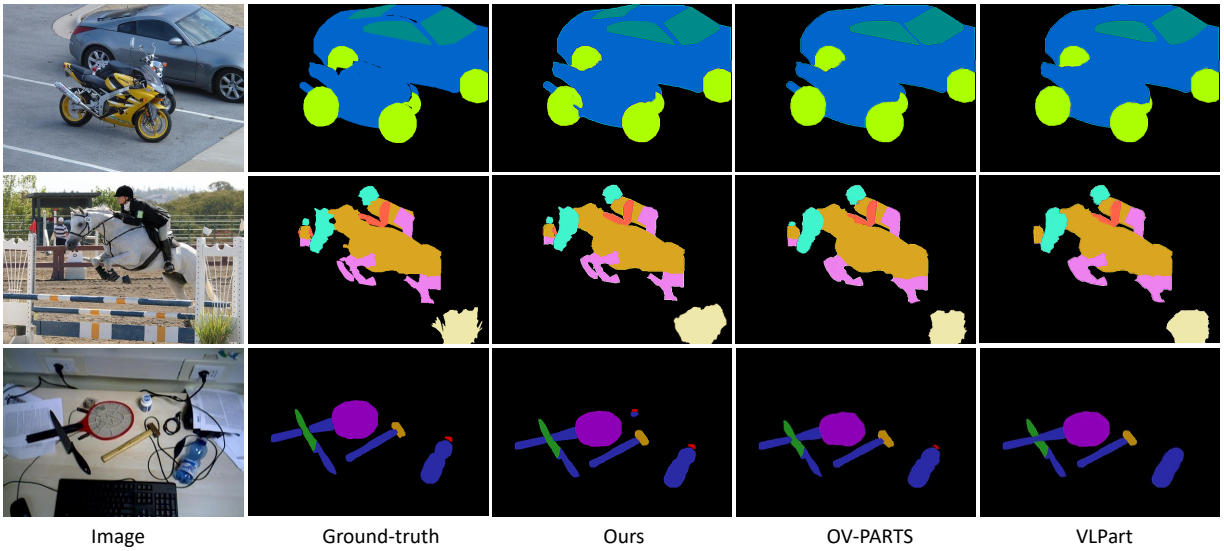


Figure 4: Qualitative comparison on proposed OVAS-25 dataset. It can be seen that our proposed method demonstrates superior performance in solving background and adjacent object pixel interference problems, as well as open-vocabulary small-size object affordance segmentation.

| Method | Param. (M) | GFLOPs | FPS |
|---|---|---|---|
| MaskFormer (Cheng et al. 2021) | 168.4 | 488.6 | 4.1 |
| VLPart (Sun et al. 2023a) | 245.2 | 548.7 | 6.4 |
| OV-PARTS (Wei et al. 2024) | 324.4 | 857.1 | 10.5 |
| Ours | **154.7** | **386.8** | **3.8** |

Table 5: Efficiency comparison with other methods.

module (MPFEM), which utilizes mask prompts to add instance features to enhance visual feature representation, and the performance improved by 6.9% mIoU. In the third row of Table 3, we continue to add fine-grained text prompts, which improves performance by 2.24% mIoU compared to only adding instance features. In the last row of Table 3, we ablate the pixel decoder. We replace the pixel decoder of MaskFormer with the proposed pixel decoder, which utilizes text prompts as knowledge prior, resulting in a performance improvement of 1.61% mIoU. These ablation experiments validate the performance of the proposed modules, and demonstrate the effectiveness of mask prompts.

Furthermore, we compare the efficiency in Table 5, i.e., trainable parameters, GFLOPs, and inference time (FPS). Since our method mainly trains parameters in the pixel decoder, the detection and segmentation modules are frozen and offline, so the parameters and GFLOPs are lowest, and the inference speed is fastest.

## Conclusion

In this paper, we propose a novel open-vocabulary affordance segmentation task and a benchmark dataset. To address the interference problems of complex background and pixels of adjacent objects, we propose an open-vocabulary affordance segmentation model that utilizes object shape mask as prompts. The model utilizes generated fine-grained masks to enhance object instance features and embed text prompts of different granularity, thereby continuously suppressing the interference and improving generalization ability in open-world. Extensive experiments on the proposed dataset verify the effectiveness of the proposed method. In addition, we also achieve competitive performance in open-vocabulary part segmentation tasks. Our findings underscore the importance of mask prompts in open-vocabulary affordance segmentation problem, and provide promising strategies for image segmentation related tasks.

## Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 62172022, GrantU21B2038, and Grant 62476179, National Key R&D Programof China No.2021ZD0111902.

## References

Balazevic, I.; Steiner, D.; Parthasarathy, N.; Arandjelović, R.; and Henaff, O. 2024. Towards in-context scene understanding. *NeurIPS*, 36.

Bastanfard, A.; Amirkhani, D.; and Mohammadi, M. 2022. Toward image super-resolution based on local regression and nonlocal means. *Multimedia Tools and Applications*, 81(16): 23473–23492.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *NeurIPS*, 32.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229. Springer.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. In *arXiv:1512.03012*.

Chen, D.; Kong, D.; Li, J.; Wang, S.; and Yin, B. 2024. ADOSMNet: a novel visual affordance detection network with object shape mask guided feature encoders. *Multimedia Tools and Applications*, 83(11): 31629–31653.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 1971–1978.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34: 17864–17875.

Cho, S.; Shin, H.; Hong, S.; An, S.; Lee, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2023. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*.

Chu, F.-J.; Xu, R.; and Vela, P. A. 2019. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4(2): 1140–1147.

Deng, S.; Xu, X.; Wu, C.; Chen, K.; and Jia, K. 2021. 3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding. In *CVPR*, 1778–1787.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *CVPR*, 11583–11592.

Do, T.-T.; Nguyen, A.; and Reid, I. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 5882–5889.

Gu, Q.; Su, J.; and Yuan, L. 2021. Visual affordance detection using an efficient attention convolutional neural network. *Neurocomputing*, 440: 36–44.

Han, K.; Liu, Y.; Liew, J. H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. 2023. Global knowledge calibration for fast open-vocabulary segmentation. In *ICCV*, 797–807.

He, Q. 2024. Prompting multi-modal image segmentation with semantic grouping. In *AAAI*, volume 38, 2094–2102.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2021. Language-driven Semantic Segmentation. In *International Conference on Learning Representations*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, S.-A.; Zhang, Y.; Qiu, Z.; Xie, H.; Zhang, Y.; and Yao, T. 2023. CARIS: Context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 779–788.

Lu, L.; Zhai, W.; Luo, H.; Kang, Y.; and Cao, Y. 2022. Phrase-Based Affordance Detection Via Cyclic Bilateral Interaction. *IEEE Transactions on Artificial Intelligence*, 1–13.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *CVPR*, 7086–7096.

Michieli, U.; Borsato, E.; Rossi, L.; and Zanuttigh, P. 2020. Gmnet: Graph matching network for large scale part semantic segmentation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 397–414. Springer.

Minh, C. N. D.; Gilani, S. Z.; Islam, S. M. S.; and Suter, D. 2020. Learning Affordance Segmentation: An Investigative Study. In *2020 Digital Image Computing: Techniques and Applications*, 1–8.

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 909–918.

Mur-Labadia, L.; Martinez-Cantin, R.; and Guerrero, J. J. 2023. Bayesian deep learning for affordance segmentation in images. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6981–6987.

Myers, A.; Teo, C. L.; Fermüller, C.; and Aloimonos, Y. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation*, 1374–1381.

Nguyen, A.; Kanoulas, D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5908–5915.

Nguyen, T.; Vu, M. N.; Vuong, A.; Nguyen, D.; Vo, T.; Le, N.; and Nguyen, A. 2023. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5692–5698. IEEE.

Pan, T.-Y.; Liu, Q.; Chao, W.-L.; and Price, B. 2023. Towards open-world segmentation of parts. In *CVPR*, 15392–15401.

Qin, J.; Wu, J.; Yan, P.; Li, M.; Yuxi, R.; Xiao, X.; Wang, Y.; Wang, R.; Wen, S.; Pan, X.; et al. 2023. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 19446–19455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ragusa, E.; Dosen, S.; Zunino, R.; and Gastaldo, P. 2023. Affordance Segmentation Using Tiny Networks for Sensing Systems in Wearable Robotic Devices. *IEEE Sensors Journal*, 23(19): 23916–23926.

Roy, A.; and Todorovic, S. 2016. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 186–201.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Sawatzky, J.; and Gall, J. 2017. Adaptive binarization for weakly supervised affordance segmentation. In *ICCV*, 1383–1391.

Singh, R.; Gupta, P.; Shenoy, P.; and Sarvadevabhatla, R. 2022. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *CVPR*, 1445–1455.

Sun, P.; Chen, S.; Zhu, C.; Xiao, F.; Luo, P.; Xie, S.; and Yan, Z. 2023a. Going denser with open-vocabulary part segmentation. In *ICCV*, 15453–15465.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2023b. Alpha-CLIP: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*.

Tan, X.; Xu, J.; Ye, Z.; Hao, J.; and Ma, L. 2021. Confident semantic ranking loss for part parsing. In *ICME*, 1–6. IEEE.

Wang, Z.; and Tian, G. 2023. Task-Oriented Robot Cognitive Manipulation Planning Using Affordance Segmentation and Logic Reasoning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Wei, M.; Yue, X.; Zhang, W.; Kong, S.; Liu, X.; and Pang, J. 2024. Ov-parts: Towards open-vocabulary part segmentation. *NeurIPS*, 36.

Wu, M.; Liu, Y.; Ji, J.; Sun, X.; and Ji, R. 2024. Toward Open-Set Human Object Interaction Detection. In *AAAI*, volume 38, 6066–6073.

Wu, Y.; Chen, J.; Yan, J.; Zhu, Y.; Chen, D. Z.; and Wu, J. 2023. GCL: Gradient-Guided Contrastive Learning for Medical Image Segmentation with Multi-Perspective Meta Labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, 463–471.

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

Xu, L.; Gao, Y.; Song, W.; and Hao, A. 2024a. Weakly Supervised Multimodal Affordance Grounding for Egocentric Images. In *AAAI*, volume 38, 6324–6332.

Xu, W.; Xu, R.; Wang, C.; Xu, S.; Guo, L.; Zhang, M.; and Zhang, X. 2024b. Spectral prompt tuning: Unveiling unseen classes for zero-shot semantic segmentation. In *AAAI*, volume 38, 6369–6377.

Yin, C.; and Zhang, Q. 2022. Object affordance detection with boundary-preserving network for robotic manipulation tasks. *Neural Computing and Applications*, 34(20): 17963–17980.

Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2024. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *AAAI*, volume 38, 6890–6898.

Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; and Zhang, L. 2023. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, 1020–1031.

Zhang, Y.; Li, H.; Ren, T.; Dou, Y.; and Li, Q. 2022. Multi-scale Fusion and Global Semantic Encoding for Affordance Detection. In *2022 International Joint Conference on Neural Networks*, 1–8.

Zhao, X.; Cao, Y.; and Kang, Y. 2020. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18): 14321–14333.

Zhao, Y.; Li, J.; Zhang, Y.; and Tian, Y. 2019. Multi-class part parsing with joint boundary-semantic awareness. In *ICCV*, 9177–9186.