Going Denser with Open-Vocabulary Part Segmentation

Peize Sun[1] Shoufa Chen[1] Chenchen Zhu[2] Fanyi Xiao[2] Ping Luo[1,4] Saining Xie[3] Zhicheng Yan[2]
[1]The University of Hong Kong [2]Meta AI [3]New York University [4]Shanghai AI Laboratory

https://github.com/facebookresearch/VLPart

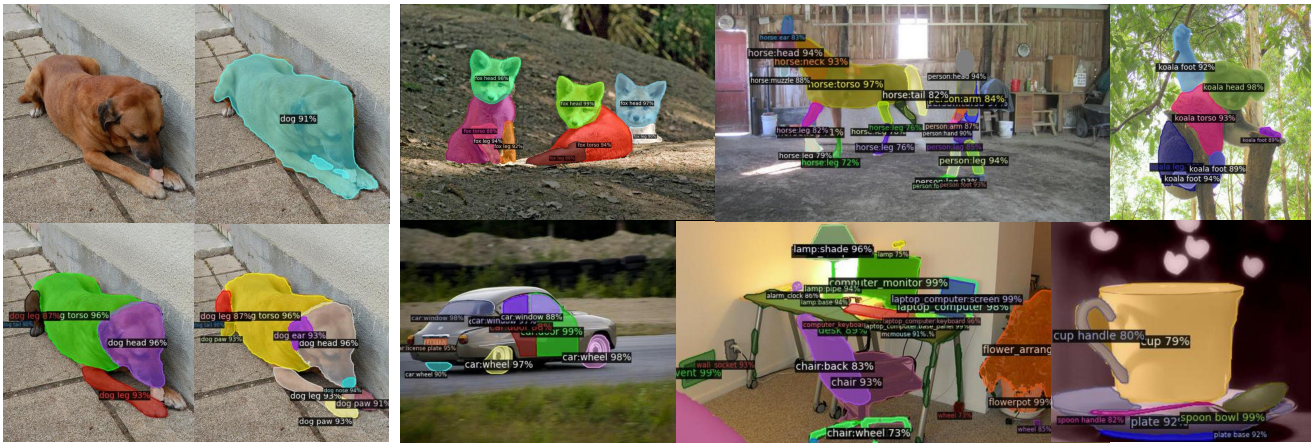

Figure 1. **Examples of open-vocabulary part segmentation.** Beyond open-vocabulary object detection, we propose that the detector should be able to predict both objects and their parts. This open-world fine-grained recognition ability is in demand for an intelligent vision system but is only realized in a limited number of categories [9, 29, 67] up to now. In this paper, we move forward to going denser with open-vocabulary part segmentation: Left figure shows segmenting `dog` and its parts in different granularities. Right figure demonstrates more visualization results.

Abstract

Object detection has been expanded from a limited number of categories to open vocabulary. Moving forward, a complete intelligent vision system requires understanding more fine-grained object descriptions, object parts. In this paper, we propose a detector with the ability to predict both open-vocabulary objects and their part segmentation. This ability comes from two designs. First, we train the detector on the joint of part-level, object-level and image-level data to build the multi-granularity alignment between language and image. Second, we parse the novel object into its parts by its dense semantic correspondence with the base object. These two designs enable the detector to largely benefit from various data sources and foundation models. In open-vocabulary part segmentation experiments, our method outperforms the baseline by 3.3∼7.3 mAP in cross-dataset generalization on PartImageNet, and improves the baseline by 7.3 novel AP_{50} in cross-category generalization on Pascal Part. Finally, we train a detector that generalizes to a wide range of part segmentation datasets while achieving better performance than dataset-specific training.

1. Introduction

Recent advances in open-vocabulary object detection [37,42,47,58,78,85,92,95] have made surprising development in enlarging the number of object categories from a pre-determined set by training datasets [17,26,49,71] to any object in the open world. This is a crucial step for the vision system to make effect in the real world. Towards the next step, for a deeper understanding to object structure, mobility, functionality, and practical applications such as behavior analysis [61,68,82], robotics manipulation [5,19,60,88], image-editing [45, 70], only object-level perception is not sufficient, while the fine-grained recognition ability of part segmentation [9, 29, 67, 83] is necessary.

Since a part is the fine-grained version of an object, an intuitive idea is to directly apply existing open-vocabulary object detection methods [42, 47, 92, 95] to solve the part detection/segmentation task. However, as shown in Table 1, they do not show good generalization on part-level recognition. Although conceptually similar to open-vocabulary object detection, localizing and classifying the fine-grained object parts are essentially more challenging. This moti-

Method	dog				
	head	leg	paw	tail	torso
RegionCLIP [92]	5.2	0.1	0.2	0.0	1.9
Detic [95]	3.2	0.0	0.0	0.0	2.0
VLDet [47]	3.5	0.0	0.0	0.0	1.9
GLIP [42]	32.6	3.1	2.7	9.5	2.2
Oracle	50.7	14.8	20.7	10.4	18.7

Table 1. **Performance of previous open-vocabulary object detection methods on Pascal Part [9] validation set.** The evaluation metric is mAP_{box}@[.5, .95] on the detailed metrics of `dog`. All models use their official codebases and model weights. Oracle is the method trained on Pascal Part training set.

vates us to explore new designs to empower current object detectors with open-vocabulary part segmentation ability.

The model of open-vocabulary part segmentation is supposed to be able to segment the object not only on open category but also on open granularity. As shown in Figure 1, the [`dog`] can be parsed to the [`head, torso, leg, tail`], while in the finer granularity, the head of a dog can be further parsed to the [`ear, eye, nose`, etc.]. Annotating such fine-grained object part is extremely expensive. Publicly available datasets of part segmentation are less rich and diverse than those of image classification and object detection datasets. Even though we collect three sources of part segmentation datasets, including Pascal Part [9], PartImageNet [29], and PACO [67], only a small number of objects part are accessible.

To expand the vocabulary of part categories, we first seek to utilize the large vocabulary object-level and image-level data, such as LVIS [26] and ImageNet [13], where object categories are known, but their part locations or part names are not. To enable part segmentation task benefit from them, our detector is based on the vision-language model [66], and trained on the joint of part-level, object-level and image-level data, where the classifier weight in the detector is replaced to the text embedding of the class name. In this way, the model learns to align vision and language at multi-granularity level to help generalize the ability to parse the object into its parts from base objects to novel objects.

Though the multi-granularity alignment is established, the part-level alignment for novel objects is fragile since its supervision signal is absent. To further strengthen it, we propose to leverage the pre-trained foundation models [7] to parse the novel object into its parts as the annotations: 1) We find the nearest base object for each novel object by the similarity of their global features. 2) We build the dense semantic correspondence between the novel object and its corresponding base object by the similarity of their spatial features. 3) We parse the novel object into its parts in the way of the base object by the correspondence. The name of novel parts follows its corresponding base object. According to this pipeline, we generate the parsed images and use them as part annotations of novel objects.

Extensive experiments demonstrate that our method can significantly improve the open-vocabulary part segmentation performance. For cross-dataset generalization on PartImageNet, our method outperforms the baseline by 3.3∼7.3 mAP. For cross-category generalization within Pascal Part, our approach improves the baseline by 7.3 AP_{50} on novel parts. Finally, we train a detector with the joint data of LVIS, ImageNet, PACO, Pascal Part, PartImageNet, and parsed ImageNet. On three trained part segmentation datasets, it obtains better performance than their dataset-specific training. Meanwhile, part segmentation on a large range of objects in the open-world is achieved , as shown in Figure 1.

Our contributions are summarized as follows:

- We set up benchmarks and baseline models for open-vocabulary part segmentation in Pascal Part and PartImageNet datasets.
- We propose a parsing pipeline to enable part segmentation to benefit from various data sources and expand the vocabulary of part categories.
- We train a detector with the ability of open-vocabulary object detection and part segmentation, achieving favorable performance on a wide range of part segmentation datasets.

2. Related Work

Open-vocabulary object detection. OVOD [85] aims to improve the generalization ability of object detectors from seen categories to novel categories. For example, ViLD [25], RegionCLIP [66], PB-OVD [18] use pseudo region annotations generated from the pre-trained vision-language model [40, 66]. DetPro [16] designs an automatic prompt learning method to improve the category embedding effectively. GLIP [42] trains the detector on both detection and grounding data. Detic [95] enlarges the number of novel classes with image classification data. VLDet [47] extracts region-word pairs from image-text pairs in an online way. Different from these works, we explore more fine-grained object recognition at the part level.

Part segmentation. Beyond recognizing objects through category labels, a more fine-grained understanding of objects at the part level [12, 43, 56, 94] is in increasing demand. Some pioneering works provide part annotations for specific domains, such as human [21, 41, 82], birds [74], cars [68, 72], fashion domain [33, 91]. Part annotations for common objects include such as Pascal-Part [9], PartNet [59], PartImageNet [29], ADE20K [93], Cityscapes-Panoptic-Parts [55] and more recent PACO [67]. Based on these valuable datasets, our work is towards parsing any object in the open world.

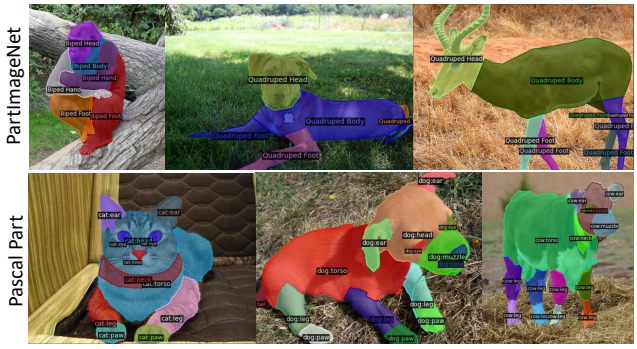

Figure 2. **Different part taxonomies in different dataset annotations.** For example, the [dog] is parsed to the [head, body, foot, tail] in PartImageNet [29]. In the finer granularity, the head of a dog is further parsed to the [ear, eye, muzzle, nose] in Pascal Part [9].

Vision-and-language representation learning. A universal representation for vision and language is needed in various tasks such as visual question answering [23, 27], image/video-text retrieval [49, 62, 79], visual reasoning [20, 73] and so on. To enhance the visual representation, region-based features obtained from object detection are introduced. For instance, OSCAR [44] uses object tags and region features to train a universal semantic, UNITER [10] establishes word-region alignments by supervising similar output across multiple modalities, ALBEF [40] aligns the image and text before fusing them with a multimodal encoder, SimVLM [76] reduces the requirement to regional labels by exploiting large-scale weak supervision. Our work is aimed at learning part-level visual representation aligned with language supervision.

Semantic correspondence. The aim of semantic correspondence [28, 34, 38, 52, 57, 80, 89] is to establish the spatial visual correspondence between different instances of the same object category. Cross-domain correspondence [1, 86] expands it to different categories. The pre-trained model is usually introduced to compute the feature map similarity. [1] used a pre-trained CNN and [3] improved the performance by using ViT model [3, 14]. In our work, we apply self-supervised DINO [7] to build the semantic correspondence between the novel object and the base object.

3. Open-Vocabulary Part Segmentation

The goal of open-vocabulary part segmentation is to parse any object *in the wild* into its components. The model takes as input the image and outputs its part segmentation by the pre-defined part taxonomy or custom text prompt.

3.1. Open Category and Granularity

We expect the model for open-vocabulary part segmentation to be able to provide the part segmentation in both open category and open granularity.

Open category. Similar to object detection, open category means that the model is able to parse any category of the object *in the wild*.

Open granularity. As shown in Figure 2, the part taxonomy is inherently hierarchical, the [dog] can be parsed to the [head, body, leg, tail], while in the finer granularity, the head of a dog can be further parsed to the [ear, eye, muzzle, nose, etc.]. This brings to inconsistent definitions in different datasets, for example, the taxonomy of Pascal Part [9] is relatively finer than ImageNetPart [29].

3.2. Evaluation Protocol

For terminology, *base parts* are the parts of base objects (seen in the training set), and *novel parts* are the parts of novel objects (unseen in the training set).

Cross-category generalization. The models are trained on the base parts, such as parts of cat, cow, horse, and sheep, and evaluated on novel categories, such as parts of dog. In this setting, the base categories and the novel categories are *mutually* exclusive.

Cross-dataset generalization. In practice use, the trained model could be used in any evaluation dataset, therefore, the object categories in inference may overlap with those in training. We use a cross-dataset setting to evaluate these more practical scenes.

3.3. Revisit Open-Vocabulary Object Detection

One may assume that the part instance is a special type of object, and the open-vocabulary part segmentation task can be solved by off-the-shelf open-vocabulary object detection/segmentation methods [42, 47, 92, 95].

Adapting open-vocabulary object detector to part recognition. To adapt the open-vocabulary object detector to part segmentation, its classifier weight in the region recognition head needs to be replaced by text embedding of the part category name. The performance of some popular open-vocabulary object detectors on part segmentation is shown in Table 1. Obviously, their performances are far from satisfactory. We analyze the limited performance of open-vocabulary object detection on part segmentation comes from two folds: *(i) Recall*. The part and the object have different granularities, it is non-trivial for region proposals trained on object-level data to generalize to part-level. *(ii) Precision*. The learning materials of open-vocabulary object detectors, object-level and image-level data, have insufficient part instances and are hard to provide effective part-level supervision.

Generalizability of region proposals. We study whether region proposal networks trained on object-level data could

Training data	Type	Pascal Part			
		AR@30	AR@100	AR@300	AR@1000
VOC	object	7.7	11.8	15.4	16.1
COCO	object	8.4	14.8	24.4	40.5
LVIS	object	12.7	20.6	30.0	45.8
Pascal Part base	part	29.4	48.1	63.6	75.3
Pascal Part	part	31.1	50.5	67.2	78.8

Table 2. **Evaluation on the generalizability of region proposals on objects and parts.** The recall is evaluated at IoU threshold 0.5 on the validation set of Pascal Part. All models are ResNet50 Mask R-CNN. The upper section is trained on object-level data and the lower section is part-level data. It is non-trivial for region proposals to generalize from object-level to part-level.

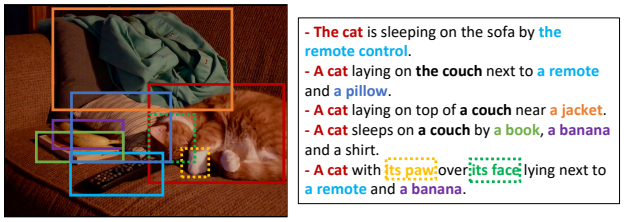

Figure 3. **Example of COCO Caption [8].** COCO Caption data provides the image and its corresponding caption only, without object-level alignment (solid box) or part-level alignment (dashed box). Even if all alignments are known, part descriptions are much less frequent than object descriptions.

provide sufficient object proposals for the part. Although previous works [25, 95] conclude that novel categories only suffer a small performance drop in recall, we point out that this is not the case when the novel objects have different granularity from object to part. As shown in Table 2, the detector trained on object-level data only has limited recall on the part dataset. To obtain better recall, part-level annotations are necessary, evidenced by the model trained on the Pascal Part base having very close recall with the fully-supervision model on the full Pascal Part training set.

Part-level alignment between image and its caption. Open-vocabulary object detection methods usually use image caption data to train the model. However, learning to detect the object part in the image from its image caption has two challenges: (1) Image caption data only provides the image and its corresponding caption, without dense captions on objects. Each open vocabulary object detector method [42, 47, 92, 95] needs to design its own method to align objects in the image and in the caption. (2) Even if the alignment could be extracted from the caption, or provided by the dataset annotations [36, 63, 84], we find the caption contains object parts less frequently than objects, as shown in Figure 3. This less frequency makes part-level alignment between the image and its caption more difficult to learn than the object-level.

4. Our Method: VLPart

Our detector architecture is a vision-language version of Mask R-CNN [30], where the classifier is the text embedding of category name from CLIP [66]. This enables us to seamlessly train the detector on part-level, object-level and image-level data. We further parse the image data into its parts to expand the vocabulary of part categories, which is based on dense semantic correspondence between the base object and the novel object extracted from DINO [7].

4.1. Detector Architecture

Image encoder. The image encoder is based on convolutional neural networks such as ResNet [31] or Transformer-based models like Swin [54], followed by Feature Pyramid Network [48] to generate multi-scale feature maps to be used in the detection decoder.

Detection decoder. The architecture of detection decoder is composed of a region proposal network (RPN) [69] and a R-CNN recognition head. RPN provides box proposals for both objects and parts. R-CNN recognition head refines the box location and the classification score. Notably, the classifier weight in the recognition head is replaced by text embedding of the class name of the object and the part.

Text embedding as the classifier. The classification score of the recognition head is implemented as a dot-product operation between the region features and the text embeddings, where the region features are cropped from feature maps of the image encoder, and the text embeddings are extracted from the text encoder in CLIP [66].

Mask decoder. We choose the architecture of mask decoder from Mask R-CNN [30] and replace the original multi-classification head with a class-agnostic head to support segmentation on novel categories. We note that more advanced architecture such as Mask2Former [11] has the potential to further improve the performance but is not the focus of this work.

4.2. Training on Parts, Objects, and Images

The training data includes part-level, object-level, and image-level data. The image data is further parsed into the part annotation. Our detector is joint-trained on these data to establish multi-granularity alignment.

Part segmentation data. Part segmentation data [9, 29, 67] contains part mask segmentation and its category. Part is always defined as an object-part pair since the same semantic part can be very different when it is associated with different objects. The category name of the part is formalized as follows:

$$C_{part} = [\text{“dog: head”, “dog: nose”, ..., “cat: tail”}]$$

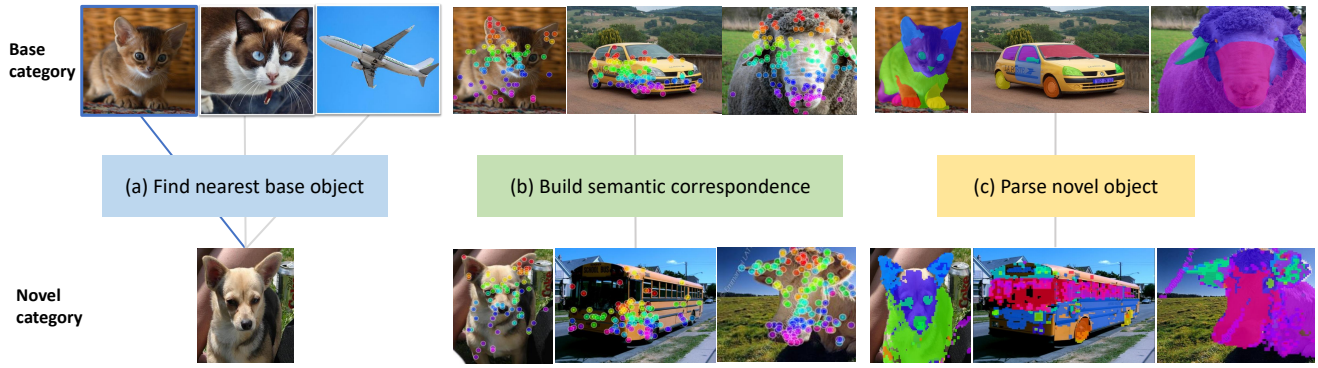

Figure 4. **The pipeline of parsing novel objects into parts.** (a) Finding the nearest base object for each novel object. (b) Building the dense semantic correspondence between a novel object and its corresponding base object. For better visualization, we only some points sampled from the feature map grid. (c) Parsing the novel object as the way of the base object.

Object detection data. Object detection data contains object boxes and its category. Most object detection datasets [26, 49] also provide object mask segmentation annotations.

$$C_{object} = [\text{“person”, “bicycle”, ..., “toothbrush”}]$$

The training loss for part and object data includes all location loss, classification loss, and mask loss.

Image classification data. Image classification data provides a large vocabulary of object categories in the form of images. Although object-level or part-level bounding annotations are absent, these images could be effectively used by the following ways: (1) The classification loss can be performed on max-size proposal [95] for each image, and therefore expands the object-level vocabulary. (2) As will be introduced in section 4.3, the image can be parsed into parts and used as part-level annotations to expand the vocabulary of part categories. The training loss about image data only includes classification loss.

4.3. Parsing Novel Objects into Parts

Most novel objects share the same part taxonomy with one of the base objects, for example, the novel `dog` has the same parts as the base `cat`. Since the part segmentation of the base object is known, we could parse the novel object according to its dense semantic correspondence to the base object. The whole pipeline is shown in Figure 4.

Finding the nearest base object for each novel object. We use DINO [7] to extract the `[class token]` of each base object, denoted as $t^{cls}(\cdot)$, and save these features as the database. Then, for each novel object i, we extract its feature using the same way and find its nearest base object i_{near} in the database by the cosine similarity.

$$i_{near} = \arg\max_{j} \ \text{sim}(t^{cls}(I_i),\ t^{cls}(I_j))$$

Building dense semantic correspondence between the base object and its nearest novel object. We further use the DINO feature map as dense visual descriptors [3], denoted as $F_{x,y}(\cdot)$, where x, y are grid indexes in the feature map. After computing the spatial similarity between the novel object $F_{x,y}(I_i)$ and its nearest base object $F_{p,q}(I_{i_{near}})$, for each token (x, y) in the novel object, its corresponding token in the base object are chosen as the token with the highest cosine similarity.

$$x_{corr}, y_{corr} = \arg\max_{p,q} \ \text{sim}(F_{x,y}(I_i),\ F_{p,q}(I_{i_{near}}))$$

Parsing novel parts by semantic correspondence. After dense correspondence between the base object and novel object is obtained, we could parse the novel object into its part segmentation $M_i(x, y)$ as the way of its corresponding base object part segmentation $M_{i_{near}}(p, q)$.

$$M_i(x, y) = M_{i_{near}}(x_{corr}, y_{corr})$$

A hybrid parser to base and novel objects. Figure 4 also provides some examples of semantic correspondence and parsed novel objects. It can be seen that the strong and clear alignment is set up and the produced part segmentation is qualified to be used as pseudo part annotation for the novel object. For the base object, we use the detector trained on base parts to generate its pseudo part annotation.

4.4. Inference on Text Prompt

In inference, the model takes as input the image and outputs the part segmentation for the object. Since all vocabulary of both objects and parts are a large number, and the user may not be interested in obtaining all possible object and part segmentation, our detector supports inference on text prompt by user input.

The left section of Figure 1 is a case using a dog as an example. When the user-input is [dog], [dog: head, torso, leg, tail] and [dog: head, ear, eye, nose, torso, leg, paw, tail], the detector outputs the segmentation results in different granularities accordingly. The right section of Figure 1 is a range of objects in the open world. It can be seen that our model is able to detect both open-vocabulary objects and their parts. When our detector is used in real applications, one can flexibly choose to use the pre-defined part taxonomy in datasets such as Pascal Part, PACO, or custom text prompt.

5. Experiment

5.1. Datasets

We use three sources of part segmentation datasets, Pascal Part [9], PartImageNet [29] and PACO [67].

Pascal Part. The original Pascal Part provides part annotations of 20 Pascal VOC classes, a total of 193 part categories. Its taxonomy contains many positional descriptors, which is not suitable for this paper, and we modify its part taxonomy into 93 part categories.

PartImageNet. PartImageNet groups 158 classes from ImageNet into 11 super-categories and provides their part annotations, a total of 40 part categories.

PACO. PACO supplements more electronic equipment, appliances, accessories, and furniture than Pascal Part and PartImageNet. PACO contains 75 object categories, 456 object-part categories and 55 attributes. The image sources of PACO are LVIS and Ego4D [24]. In this work, we use PACO-LVIS set as default. We focus on object parts and leave attributes for future research.

For object-level detection data, we use VOC [17], COCO [49] and LVIS [26]. For image-level data, we use ImageNet1k (**IN**) [13]. We also create ImageNet-super11 (**IN-S11**) and ImageNet-super20 (**IN-S20**) that overlap with PartImageNet and Pascal category vocabulary separately. More details about datasets are in Appendix.

5.2. Cross-dataset segmentation on PartImageNet

In Table 3, we study cross-dataset generalization by using PartImageNet validation set as the evaluation dataset, where the metrics of all (40) parts and the detailed metrics of parts of quadruped are reported.

Table 3a shows when Pascal Part is the only available human-annotated part dataset, using IN-S11 data could help to improve PartImageNet performance.

Baseline from Pascal Part. The baseline method directly uses the Pascal Part-trained model to evaluate PartImageNet. As shown in Table 3a first row, the performance is poor, for example, body and foot of the quadruped are nearly to zero. Pascal Part has no semantic label of quadruped, and the model needs to generalize from parts of dog, cat, etc. in Pascal Part to parts of quadruped in PartImageNet. The possible generalization ability comes from the text embedding generated from CLIP [66]. However, generalization in part-level recognition is beyond its capability since CLIP is pre-trained on only image-level data.

Method	All (40)	quadruped			
		head	body	foot	tail
Pascal Part	4.5	17.4	0.1	0.0	2.9
+ IN-S11 label	5.4	23.6	3.4	0.8	1.2
+ Parsed IN-S11	7.8	35.0	15.2	3.5	8.9
vs. baseline	**+3.3**	**+17.6**	**+15.1**	**+3.5**	**+6.0**
PartImageNet	29.7	57.3	25.8	22.9	22.9

(a) **Cross-dataset generalization when only one part dataset, Pascal Part, is available.** Pascal Part is trained on the Pascal Part training set. IN-S11 label and Parsed IN-S11 are added into the training sequentially.

Method	All (40)	quadruped			
		head	body	foot	tail
Pascal Part	4.5	17.4	0.1	0.0	2.9
+ LVIS, PACO	7.8	22.9	7.1	0.3	4.0
+ IN-S11 label	8.8	26.3	3.7	0.4	1.0
+ Parsed IN-S11	11.8	47.5	13.4	4.5	14.8
vs. baseline	**+7.3**	**+30.1**	**+13.3**	**+4.5**	**+11.9**
PartImageNet	29.7	57.3	25.8	22.9	22.9

(b) **Cross-dataset generalization when more than one part datasets are available.** Starting from Pascal Part, LVIS, PACO, IN-S11 and Parsed IN-S11 are added into the training sequentially.

Table 3. **Cross-dataset generalization on PartImageNet part segmentation.** The evaluation metric is mAP_{mask}@[.5, .95] on the validation set of PartImageNet. All models are ResNet50 Mask R-CNN and use the text embedding of the category name as the classifier. PartImageNet is the fully-supervised method as the oracle performance.

IN-S11 label. Considering that Pascal Part has no semantic label such as quadruped, piped, *etc.*, we collect IN-S11 images from ImageNet and add them to the training as image-level classification data. As shown in Table 3a second row, the performance is improved to some extent. This shows that image-level alignment is beneficial to the part recognition task. However, since no additional part-level supervision signal is introduced when using IN-S11 as image classification data, the improvement is still limited.

Parsed IN-S11. We use our parsing pipeline to deal with IN-S11 images and generate their part annotations. As shown in the third row in Table 3a, introducing these parsed parts into the training brings a significant improvement, 3.5∼17.6 mAP improvement on the parts of quadruped and 3.3 mAP gain on all 40 parts over the baseline method.

Method	All (93)		Base (77)		Novel (16)	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
Base part	15.0	33.4	17.8	39.6	1.5	3.7
+ VOC object	16.8	36.8	19.9	43.3	2.1	5.9
+ IN-S20 label	17.4	37.5	20.8	44.7	1.1	3.1
+ Parsed IN-S20	18.4	39.4	21.3	45.3	4.2	11.0
vs. baseline	**+3.4**	**+6.0**	**+3.5**	**+5.7**	**+2.7**	**+7.3**
Pascal Part	19.4	42.7	18.8	41.5	22.1	48.9

Table 4. **Cross-category generalization on Pascal Part part segmentation**. The evaluation metric is on the validation set of the Pascal Part. All models are ResNet50 Mask R-CNN and use the text embedding of the category name as the classifier. Base part is the base split from Pascal Part. VOC object, IN-S20 label and Parsed IN-S20 are added into the training sequentially. Pascal Part is the fully-supervised method as the oracle performance.

This suggests that our proposed methods are able to provide an effective part-level supervision signal to the detection model and boosts its performance on cross-dataset generalization.

More part datasets are available. Table 3b shows when more than one human-annotated part datasets are available, including Pascal Part, PACO, and LVIS. Although LVIS is an object-level dataset, we find its categories contain many object parts, such as shoes, which can also be seen as parts. From the first two rows of Table 3b, we can see that when the part-level annotations grow in training, the part segmentation obtains better performance, from 4.5 mAP to 7.8 mAP. When IN-S11 label and parsed IN-S11 are added to the training, the performance is further boosted by a large margin. For example, the head of `quadruped` has achieved 47.5 mAP, close to fully-supervised 57.3 mAP. This shows that when more data sources are available in the future, a strong model for part segmentation in the open world is promising.

5.3. Cross-category segmentation on Pascal Part

We evaluate the cross-category generalization within the Pascal Part dataset. All 93 parts are split into 77 base parts and 16 novel parts, detailed in Appendix. Table 4 reports the metrics of all (93), base (77), and novel (16) parts.

Baseline from Pascal Part base. Table 4 first row is the baseline, which is trained on base parts and evaluated on novel parts. Since the detector uses CLIP text embedding as the classifier, the novel parts obtain non-zero segmentation performance.

VOC object. Compared with the part annotation, the object annotation is much easier to collect. We add VOC object data to verify whether this could help to improve the performance. As shown in the second row of Table 4, adding VOC object data helps to improve the performance on both base parts and novel parts in Pascal Part. This demonstrates that object-level alignment could lead to better part-level performance.

Method	PartImageNet		Pascal Part		PACO	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
Joint	29.1	52.0	22.6	47.8	9.3	18.9
+ IN	30.8	54.4	23.6	49.2	9.0	18.7
+ Parsed IN	31.6	55.7	24.0	49.8	9.6	20.2
vs. baseline	**+2.5**	**+3.7**	**+1.4**	**+2.0**	**+0.3**	**+1.3**
Dataset-specific	29.7	54.1	19.4	42.3	10.6	21.7

(a) All models are ResNet50 [31] Mask R-CNN [30].

Method	PartImageNet		Pascal Part		PACO	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
Joint	40.0	64.8	31.2	60.5	15.4	30.3
+ IN	41.2	66.8	31.7	61.1	15.9	30.8
+ Parsed IN	42.0	68.2	31.9	61.6	15.6	30.6
vs. baseline	**+2.0**	**+3.4**	**+0.7**	**+0.9**	**+0.2**	**+0.3**
Dataset-specific	41.7	68.7	27.4	56.1	15.2	29.4

(b) All models are Swin-B [54] Cascade Mask R-CNN [6].

Table 5. **Part segmentation across datasets.** All models are evaluated by setting the classifier as text embedding of category name in the evaluation dataset. Joint denotes the joint-training on LVIS, PartImageNet, Pascal Part and PACO datasets. Dataset-specific uses the training data of each dataset, separately.

IN-S20 label. Image-level classification data is also an easy-to-get annotation. We collect images with Pascal categories from ImageNet, IN-S20, and add them to the training. As shown in Table 4 third row, additional image-level data does not bring much gain than object detection data. This is because image-level data has a similar effect as object-level data on part-level recognition. Most of its gain is diminished by object data.

Parsed IN-S20. We use our proposed parsing method to generate part annotations for novel objects, and they provide supervision on part classification. As shown in Table 4 fourth row, our method improves the performance on both base and novel categories. This shows that our parsing pipeline is an effective solution to both base and novel object part segmentation.

5.4. Part segmentation across datasets

Towards detecting and parse *any* object in the open world, we train a detector on the joint of available part segmentation datasets, including LVIS, PACO, Pascal Part and PartImageNet. The performance is shown in Table 5.

This joint training model shows good generalization ability on various evaluation datasets, for example, Pascal Part obtains 22.6 mAP, better performance than its dataset-specific training. However, the potential problem lies in that

Pascal Part	All (93)	dog head	dog torso	dog paw	dog tail
a [object] [part]	19.1	50.7	18.7	20.7	10.4
[part] of a [object]	18.4	48.8	17.6	21.3	9.2
PartImageNet	All (40)	quadruped head	quadruped body	quadruped foot	quadruped tail
a [object] [part]	29.7	57.3	25.8	22.9	22.9
[part] of a [object]	29.9	55.9	25.1	22.9	24.3

Table 6. **Text prompt template to object part.** We compare different templates of text prompt to object part in the fully-supervision setting of Pascal Part and PartImageNet.

Method	All (40)	quadruped head	quadruped body	quadruped foot	quadruped tail
Baseline	5.4	23.6	3.4	0.8	1.2
Max-score [92]	6.0	29.6	7.1	1.0	1.7
Max-size [95]	5.3	20.5	3.5	0.6	4.7
Parsed (ours)	7.8	35.0	15.2	3.5	8.9

Table 7. **Comparisons of different aligning methods for novel parts.** The experiments are carried out on cross-dataset generalization from Pascal Part to PartImageNet. Fine-tuning from the baseline model, max-score, max-size and our method apply different designs to utilize image-level data to further improve part segmentation performance, where the former two are trained on part labels expanded from the image label.

joint training does not benefit all datasets, where PartImageNet and PACO decrease the performance a little.

To make up for the performance loss, we add IN and Parsed IN into the training. It can be seen all datasets obtain the performance gain accordingly. When we scale up the model capability from ResNet50 [31] to Swin-B [54], the detector achieves better performance than dataset-specific training on all Pascal Part, PartImageNet and PACO datasets.

5.5. Ablation Study

Text prompt template. Since the part is associated with the object category, we study how to design the text prompt template of (`object`, `part`) pair to the text encoder. We select two common expressions: a [`object`] [`part`] and [`part`] of a [`object`]. For example, [`dog`] and [`head`], these two expressions are [`a dog head`] and [`head of a dog`]. As shown in Table 6, a [`object`] [`part`] behaves a little better than [`part`] of a [`object`] in Pascal Part while not in PartImageNet. Which expression is a generally better usage of text prompt to the part needs to be verified on more datasets and we leave it for future research. In addition, more advanced prompt engineering for part segmentation is also an open problem.

Figure 5. **Semantic correspondence from vanilla ViT and DINO.** The upper section is from supervised ViT model [14] and the lower section is from self-supervised DINO [7]. For each section, first row and second row are paired base objects and novel objects. We crop each image into a uniform size for better visualization.

Aligning method for novel parts. We compare different aligning methods to use IN-S11 data to help part segmentation in PartImageNet. We select two popular designs in open-vocabulary object detection, max-score and max-size. Max-score is selecting the proposal that has the highest score of the target category as the matched proposal, used in [92]. Max-size is selecting the proposal that has the maximum area among all proposals as the matched proposal to the target category, proposed in [95]. For each ImageNet image, its object category is known, and its part taxonomy can be inferred, these parts will be used as the target category in max-score and max-size methods.

- *Max-score.* As shown in Table 7 second row, max-score helps to improve the performance a little over baseline. Fine-tuning from the baseline model, its selected highest score proposals contain efficient training samples, and these samples bring performance gain.

- *Max-size.* As shown in Table 7 third row, the max-size method degenerates the performance in most metrics. According to the max-size rule, all parts are assigned to the same proposal, it is hard to align the part-level region to its part name text embedding. This shows that part-level alignment is more difficult than object-level and an efficient fine-grained supervision signal is necessary.

Pre-trained model in semantic correspondence. We use self-supervised DINO [7] in this work to find the nearest base object for each novel object and build their dense

Source	capability	part name	part location
base parts	Align image and text in part-level of base objects	✓	✓
novel objects	Align image and text in object-level and image-level of novel objects	✓	
CLIP [66]	Anchor the part name in language feature space	✓	
DINO [7]	Parse the novel object into its part		✓

Table 8. **The source of VLPart capability.** Besides training data of base parts and novel objects, two foundation models, CLIP and DINO, contribute to open-vocabulary part segmentation.

semantic correspondence. We verify whether a vanilla ViT [14] has a similar function, which is pre-trained on fully-supervised ImageNet. As shown in Figure 5, vanilla ViT is obviously behind DINO in the aspect of providing semantic correspondence. On the one hand, the nearest base object found by DINO has better alignment with the novel object in color, texture, and pose. On the other hand, the dense correspondence from DINO has clear semantics correspondence between the two objects. Similar experiment phenomenons are reported in [3, 7]. Besides DINO, whether other models could benefit to part segmentation is a potential research direction in the future.

6. Discussion

Learning from Foundation Models. When we analyze how VLPart achieves open-vocabulary part segmentation capability, as shown in Table 8, we could see that the important components of VLPart's capability are two foundation models: CLIP [66] and DINO [7]. Learning from foundation models is a recently rising research topic [2, 4, 15, 39]. Although a single foundation model is not an expert in a specific task, for example, neither CLIP nor DINO can accomplish the part segmentation task, combining these foundation models could bring to a range of applications [22, 32, 46, 50, 53, 77, 87, 90], and this paper takes the part segmentation task as an example to explore. In the future, how to "decode" more capabilities from foundation models is a very promising research topic.

Comparison with Segment Anything Model. Segment Anything Model (SAM) [35] is a recently proposed model aimed to generate masks for all entities [64, 65] in an image, including both objects and their parts. As shown in Figure 6, the main differences between SAM and VLPart are: (1) SAM is a class-agnostic mask segmentation model, while VLPart is class-aware. (2) The part segmentation of SAM is mostly edge-oriented, which makes it hard to parse two parts if there is no obvious edge between them, while VLPart parses objects based on semantics instead of low-level edge signals.

Figure 6. **Comparsion of SAM [35] and VLPart.** The main differences are: (1) SAM is a class-agnostic segmentation model and VLPart is class-aware, (2) SAM parses the object mostly in an edge-oriented way and VLPart is semantic-oriented.

Segment and Recognize Anything Model. A big picture for the vision perception system is to segment and recognize anything in the open world, in which SAM, open-vocabulary object detection and our open-vocabulary part segmentation are all sub-tasks. Some recently public works [51, 75, 81, 96] attempt to achieve this goal but their focuses are either segmentation or recognition. Furthermore, their explorations only reach the object-level, and do not go denser into the part-level. This paper provides a promising solution to part-level segmentation and recognition, serving as a component of achieving the goal of Segment and Recognize Anything Model.

7. Conclusion and Future Work

In this paper, we explore to enable object detectors with the fine-grained recognition ability of open-vocabulary part segmentation. Our model a vision-language version of the segmentation model to support text prompt input. The training data is the joint of part-level, object-level and image-level data to establish multi-granularity alignment. To further improve the part recognition ability, we parse the novel object into its parts by the dense semantic correspondence with its nearest base objects. Extensive experiments show that our method can significantly improve the open-vocabulary part segmentation performance and achieve favorable performance on a wide range of datasets.

In the future, our models have great potential to be applied to various applications such as robotic manipulation [19], part-guided instance object [67], and part-aware image editing [45].

Acknowledgments. This work was done when Peize Sun worked as an intern at Meta AI and was supported in part by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 9

[3] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 3, 5, 9

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 9

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 7

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 5, 8, 9

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[9] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 1, 2, 3, 4, 6

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 3

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4

[12] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 8, 9

[15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 9

[16] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 6

[18] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards open vocabulary object detection without human-provided bounding boxes. *arXiv preprint arXiv:2111.09452*, 2021. 2

[19] Yuying Ge, Annabella Macaluso, Li Erran Li, Ping Luo, and Xiaolong Wang. Self-play and self-describe: Policy adaptation with vision-language foundation models. *arXiv preprint arXiv:2212.07398*, 2022. 1, 9

[20] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019. 3

[21] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. 2

[22] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 9

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 6

[25] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language

knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 4

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 2, 5, 6

[27] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[28] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE international conference on computer vision*, pages 1831–1840, 2017. 3

[29] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. *arXiv preprint arXiv:2112.00933*, 2021. 1, 2, 3, 4, 6

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 7

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7, 8

[32] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 9

[33] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 316–332. Springer, 2020. 2

[34] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6560–6569, 2017. 3

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 9

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[37] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 1

[38] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 3

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 9

[40] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3

[41] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 2

[42] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2, 3, 4

[43] Xiangtai Li, Shilin Xu, Yibo Yang Cheng, Yunhai Tong, Dacheng Tao, et al. Panoptic-partformer: Learning a unified model for panoptic part segmentation. *ECCV*, 2022. 2

[44] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[45] Yuheng Li, Krishna Kumar Singh, Yang Xue, and Yong Jae Lee. Partgan: Weakly-supervised part decomposition for image generation and segmentation. In *British Machine Vision Conference (BMVC)*, 2021. 1, 9

[46] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023. 9

[47] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 1, 2, 3, 4

[48] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence

Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 3, 5, 6

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 9

[51] 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. 9

[52] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 3

[53] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language, 2023. 9

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 7, 8

[55] Panagiotis Meletis, Xiaoxiao Wen, Chenyang Lu, Daan de Geus, and Gijs Dubbelman. Cityscapes-panoptic-parts and pascal-panoptic-parts datasets for scene understanding. *arXiv preprint arXiv:2004.07944*, 2020. 2

[56] Umberto Michieli, Edoardo Borsato, Luca Rossi, and Pietro Zanuttigh. 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*, pages 397–414. Springer, 2020. 2

[57] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 3

[58] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 1

[59] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 2

[60] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 1

[61] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 1

[62] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3

[63] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 4

[64] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. Fine-grained entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 9

[65] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 9

[66] 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, 4, 6, 9

[67] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023. 1, 2, 4, 6, 9

[68] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1906–1915, 2018. 1, 2

[69] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[71] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[72] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5452–5462, 2019. 2

[73] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 3

[74] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *technical report*, 2011. 2

[75] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 9

[76] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[77] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 9

[78] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. *arXiv preprint arXiv:2302.13996*, 2023. 1

[79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[80] Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2785, 2017. 3

[81] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 9

[82] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 364–373, 2019. 1, 2

[83] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 1

[84] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 4

[85] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2

[86] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 3

[87] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023. 9

[88] Rufeng Zhang, Tao Kong, Weihao Wang, Xuan Han, and Mingyu You. 3d part assembly generation with instance encoded transformer. *IEEE Robotics and Automation Letters*, 7(4):9051–9058, 2022. 1

[89] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3354–3364, 2021. 3

[90] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. *arXiv preprint arXiv:2212.04501*, 2022. 9

[91] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678, 2018. 2

[92] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. *arXiv preprint arXiv:2112.09106*, 2021. 1, 2, 3, 4, 8

[93] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2

[94] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1622–1631, 2021. 2

[95] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 1, 2, 3, 4, 5, 8

[96] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 9