PartGLEE: A Foundation Model for Recognizing and Parsing Any Objects

Junyi Li[1*], Junfeng Wu[1*], Weizhi Zhao[1], Song Bai[2], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology
[2] ByteDance Inc.

Abstract. We present PartGLEE, a part-level foundation model for locating and identifying both objects and parts in images. Through a unified framework, PartGLEE accomplishes detection, segmentation, and grounding of instances at any granularity in the open world scenario. Specifically, we propose a Q-Former to construct the hierarchical relationship between objects and parts, parsing every object into corresponding semantic parts. By incorporating a large amount of object-level data, the hierarchical relationships can be extended, enabling PartGLEE to recognize a rich variety of parts. We conduct comprehensive studies to validate the effectiveness of our method, PartGLEE achieves the state-of-the-art performance across various part-level tasks and obtain competitive results on object-level tasks. The proposed PartGLEE significantly enhances hierarchical modeling capabilities and part-level perception over our previous GLEE model. Further analysis indicates that the hierarchical cognitive ability of PartGLEE is able to facilitate a detailed comprehension in images for mLLMs. The model and code will be released at `https://provencestar.github.io/PartGLEE-Vision/`.

Keywords: Foundation Model · Hierarchical Recognition · Part Segmentation

1 Introduction

In recent years, foundation models have dominated the majority of tasks in the fields of Natural Language Processing [3, 9, 54] and Computer Vision [19, 23, 53, 60, 71, 72]. CLIP family [12, 13, 21, 53, 81] have made significant advancements in transfer learning and have demonstrated impressive zero-shot capabilities on vision-language tasks. SAM [23] has revolutionized the development of segmentation tasks and is able to provide multi-level class-agnostic masks. GLEE [71] utilized diverse object-level data to develop general object representations, enabling detection, segmentation, tracking, grounding, and identification of objects in open-world scenarios. Their remarkable achievement can be attributed to the integration of extensive and diverse range of datasets.

[*] Equal Technical Contribution. Work done during Junfeng's internship at ByteDance.
[†]Correspondence to Xiang Bai <`xbai@hust.edu.cn`>.

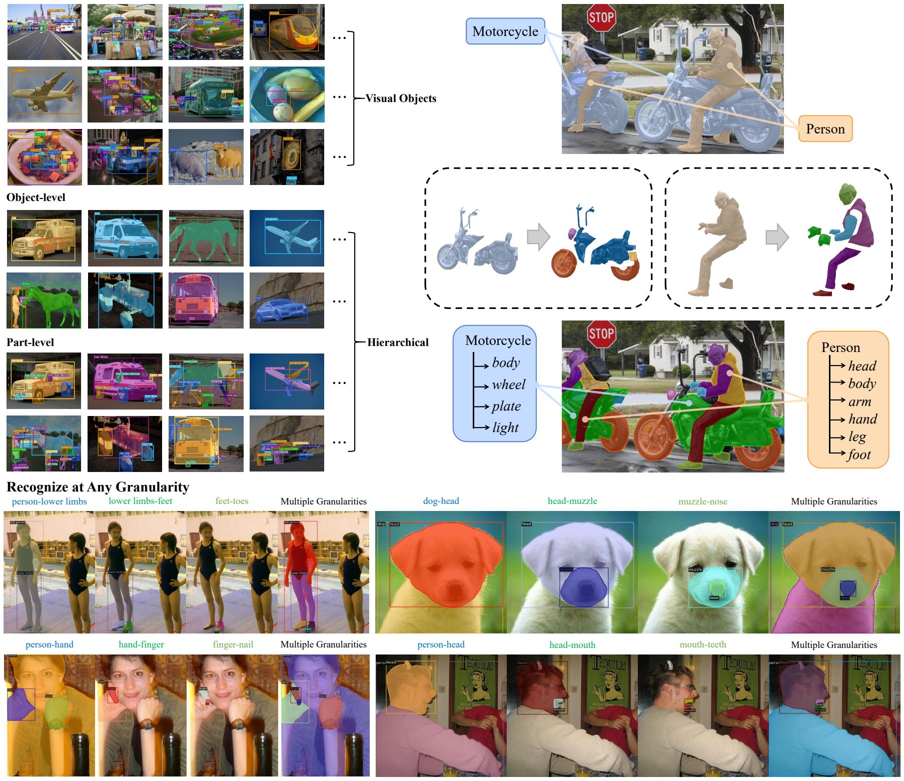

Fig. 1: An illustrative example demonstrating image annotations at diverse granularities across multiple datasets. The annotations at hierarchical levels with corresponding relationships are depicted on the right side. Below is a visualization of our segmentation results at multiple granularities.

Different from the vast quantity of object-level data, the scale of part-level data is relatively small, which turns out to be a major bottleneck hindering vision models from recognizing part-level instances. Thus, most vision models lack the hierarchical comprehension between objects and parts. However, it is evident that the ability to recognize parts from objects is essential for various practical applications such as image editing [22,33,38], behavior analysis [50,75], pose estimation [10,76], robotics manipulation [2,49], etc. Moreover, we observe that part-level information is able to help multi-modal Large Language Models (mLLMs) in achieving a more detailed understanding of image content. Since part-level comprehension is a critical ability for foundation models to tackle a broader range of problems, it leads to a natural question: How could we break through data limitations to build a part-level vision foundation model?

To enable object foundation model with part-level cognitive ability, we emphasize that the model should achieve two key objectives: (1) **Hierarchical**

Comprehension, the model is supposed to understand the intrinsic relationship between objects and parts, and extend this hierarchical connection to any novel object, (2) **Semantic Granularity**, the model should be capable of learning a universal feature representation, enabling it to recognize semantic instances at any granularity. Consequently, we present a method to jointly detect and segment both objects and parts in a top-down manner. A lightweight Querying Transformer (Q-Former) is proposed to construct the hierarchical relationship between objects and parts. Specifically, it employs a set of universal parsing queries to interact with object queries, consequently generating multiple part-level queries that are capable of predicting corresponding semantic parts for each object. The Q-Former acts as a decomposer, which first recognizes individual objects in the images and subsequently parsing them into parts. Such model design is built upon the observation that various common objects often exhibit shared characteristics of parts. For example, cats, dogs, and dinosaurs all have parts such as torso, legs, and tails. In this way, two sets of query embeddings at different levels are generated, which are then used to predict object-level and part-level instances respectively. Through this approach, the relationship between objects and parts is established via the Q-Former design. Meanwhile, the hierarchical levels of objects and parts are distinguished, which is different from previous research [6,55,62,68,70] that consider parts as fine-grained objects. This paradigm enables vision models to better understand the features on different levels during training, thereby achieving improved performance.

Our complete solution, PartGLEE, for jointly detecting and segmenting instances at both object and part levels, makes it possible for vision models to achieve favorable outcomes on both object and part levels. Some previous research have devised specialized training paradigms to utilize abundant image-text pair data [79, 89] as well as grounding data [30, 39, 71], thereby enhancing the cognitive and generalization capabilities of the models. On the contrary, the quantity of part-level data is much smaller compared to object-level data. So far, the largest dataset incorporating the concepts of both objects and parts is the recently proposed PACO [55] dataset. The scarcity of data has limited research on part-level recognition and restricted the generalization improvement of vision models. Although VLPart [62] has attempted to utilize pseudo-labeling schemes to generate part-level annotations for both object-level and image-level datasets, the quality of the pseudo-labels is relatively poor. Our innovative algorithm that parsing objects into their corresponding parts facilitates the transfer of generalization capability from objects to parts. Consequently, parts are generated from objects, which enables vision models to maintain generalization performance when predicting parts for novel objects without labeling extensive part-level data. To facilitate the training process of Q-Former, we standardize the annotation granularity across various part-level datasets and introduce a vast amount of object-level datasets, an intuitive display of the overall training data is shown in Fig. 1. Unlike VLPart, which exhibits unsatisfactory performance at object-level datasets after joint-training, our method demonstrates favorable outcomes at both object and part levels after joint-training. Moreover, it turns out that using object-level datasets is able to improve the performance of the model on part-level tasks, indicating a beneficial interaction between objects and parts.

Extensive experiments demonstrate that our method significantly improve the open-vocabulary part segmentation performance, concurrently ensuring a decent performance on object detection and segmentation. We verify its effectiveness on various popular datasets. To validate the generalization performance of our model in identifying various parts of novel objects, we conduct experiments on PartImageNet [18] and Pascal Part [6] datasets in cross-dataset and cross-category manners respectively. Our method exhibits strong transferability and generalization ability when adding extra object-level datasets during training. To evaluate the decomposition capability of our model, we conduct experiments on both ADE20K-Part and Pascal Part datasets follow OV-PARTS [70]. As a result, our approach significantly outperforms one-stage baselines of OV-PARTS, with an increase of 8.16% and 2.07% on harmonic mean IoU (hIoU) in ADE20K-Part-234 and Pascal-Part-116 respectively. Additionally, by incorporating a large amount of object-level data for joint-training, our method establishes generic hierarchical relationships and breaks through the limitations of scarce part-level data, achieving state-of-the-art performance across various part-level tasks.

In conclusion, our main contributions can be summarized as follows:

1. We construct the hierarchical relationship between objects and parts via the Q-Former, facilitating part segmentation to acquire advantages from various object-level datasets.
2. We propose a unified pipeline for hierarchical detection and segmentation, where we first recognize objects and then parsing them into corresponding semantic parts. This algorithm enables us to jointly detect and segment both object-level and part-level instances.
3. We standardize the annotation granularity across various part-level datasets by incorporating corresponding object-level annotations, complementing the hierarchical correspondences for current part-level datasets, promoting the development of vision foundation models.

2 Related Work

2.1 Visual Foundation Models and General Models

Vision foundation models and generalist models are considered as a milestone in the development of the intelligent vision system. For instance, multi-modal visual foundation models [1, 21, 53, 67, 81] have significantly advanced efficient transfer learning and exhibit impressive zero-shot capabilities on vision-language tasks by using contrastive learning with large-scale image-text pairs. Generative foundation models [11, 56, 57, 60] are trained on vast collections of images and captions, empowering them to generate image content conditioned on textual prompts. Self-supervised foundation models [4, 12, 13, 19] have learned general visual representations from large-scale image datasets, enhancing their ability to transfer to downstream tasks. However, the image-level features learned by these foundation models are not well-suited for direct application to dense prediction tasks that involve precise object and part localization.

Transformer-based generalist methods [5, 41, 66, 77, 87] adopt a sequence generation pipeline to unify the output of text and spatial coordinates. However, they mainly focus on image-level comprehension, which results in relatively weak localization capabilities. Works such as UNINEXT, etc. [28, 73, 74], built upon strong detectors [27, 86], demonstrating a strong localization capability across multiple datasets. But they fail to exhibit zero-shot transfer ability and generalization capability due to their closed-set training paradigm. Some works about open-vocabulary detection (OVD) [30, 34, 35, 42, 43, 78, 82, 83] have explored zero-shot generalization capabilities on novel categories. X-Decoder [89] and SEEM [90] have developed a versatile decoding architecture that are able to generate accurate pixel-level segmentation predictions. GLEE [71] addresses various object-level tasks through a unified architecture and training paradigm. However, current generalist models and foundation models are trained mainly on image-level and object-level datasets, thus their ability to extract more fine-grained information is limited, making it difficult for them to recognize corresponding semantic parts of any object. Our work focuses on empowering hierarchical cognitive capability for vision foundation models, thereby further advancing the development of comprehensive visual systems.

2.2 Part Segmentation

The growing interest in achieving a more fine-grained understanding of objects has sparked a surge in research focused on part level recognition. Some pioneering studies have introduced datasets with part-level annotations, concentrating on objects of some specific categories such as human body parts [15, 29, 75], animal body parts [64] and vehicle components [58]. More general part annotations for common objects such as Pascal-Part [6], PartImageNet [18], ADE20K [84], CityscapesPanoptic-Parts [45] and more recent PACO [55] were then proposed to promote more in-depth research in the field of parts. Most of the previous works [14, 31, 46, 85] were conducted based on a closed-set configuration, thus only capable of detecting and segmenting closed-set objects and parts. Recently, VLPart [62] present a pipeline for detecting and segmenting both open-vocabulary objects and their corresponding part regions, while OV-PARTS [70] utilize adapters to transfer the generalization abilities of CLIP into open-vocabulary part segmentation task. However, due to the limited quantity of data, the generalization capability of previous models [15, 29, 51, 62, 63, 68, 70, 75] still relies heavily on the training datasets. Furthermore, in prior works, both objects and parts are treated equally, they consider part as a special type of object. On the contrary, we distinguish them by considering parts as integral components subordinate to objects and generate parts from corresponding objects in a top-down manner. Our work is aimed at building hierarchical relationships while unifying the training paradigm for object and part-level data. By incorporating a large amount of object-level data, the hierarchical relationships can be extended to any object, enabling our method to recognize a rich variety of parts.

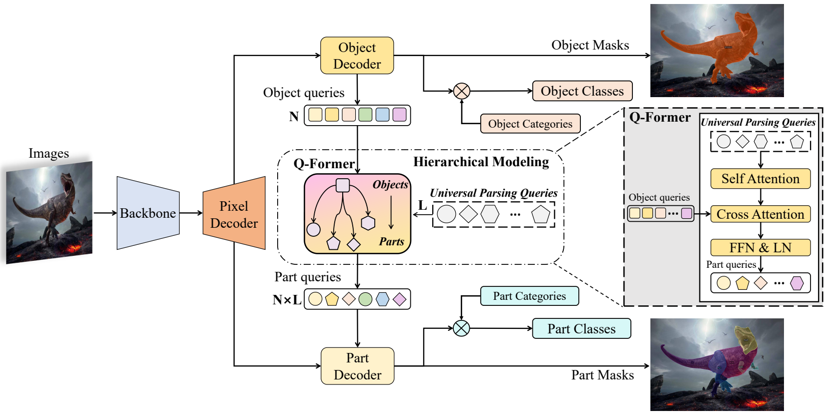

Fig. 2: Framework of PartGLEE. The Q-Former takes each object query as input and output the corresponding part queries. These queries are then fed into the object decoder and the part decoder respectively to generate hierarchical predictions.

2.3 Hierarchical Learning of Objects and Parts

Learning objects through parts has been a long-standing research topic as part annotations provide more detailed semantic information of objects. Morabia et al. [48] first introduced a pipeline employing an attention mechanism for simultaneous detection of both objects and parts. Deepflux [69] designed an image context flux representation which enables better object parts interaction for skeleton detection. Leopart [88] demonstrated that learning object parts can provide spatially diverse representation which facilitates self-supervised semantic segmentation. Wang et al. [65] proposed a method to predict both parts and objects simultaneously on Pascal-Part dataset [6]. Recent studies such as SAM [23] and Semantic-SAM [26] have studied on class-agnostic multi-granularity interactive segmentation task. However, they have not explored the relationship between objects and their corresponding semantic parts. Recently, Compositor [17] designed a bottom-up pipeline to predict parts and then cluster them into objects, while AIMS [52] utilized an independent relation decoder to construct the hierarchical association between objects and parts. Different from these works, our approach introduces a Querying Transformer to model the hierarchical relationship, allowing our model to parse any object into its corresponding parts.

3 Method

3.1 Overall Framework

Following [68, 71], we propose PartGLEE, which comprises of an image encoder, a Q-Former, two independent decoders and a text encoder, as shown in Fig. 2.

Given an input image $I \in \mathcal{R}^{H \times W \times 3}$, the backbone and the pixel decoder first extract multi-scale image features $F_s \in \mathcal{R}^{\frac{H}{2^s} \times \frac{W}{2^s} \times C}$ and $s = \{2, 3, 4, 5\}$ with backbones such as ResNet [20] or Swin Transformer [40]. Then we feed them into the object decoder, where the object-level query embeddings $q_{obj} \in \mathcal{R}^{N \times C}$ are generated in a two-stage process. These object queries are utilized to perform object-level classification, detection as well as segmentation tasks through three independent prediction heads. Besides, the object queries q_{obj} are fed into the Q-Former simultaneously, where L learnable universal parsing queries are initialized to interact with object queries. It takes object queries as input and generate part-level queries $q_{part} \in \mathcal{R}^{N \cdot L \times C}$ which are then passed into the part decoder to yield part-level predictions (detailed in Sec. 3.2). To enhance the semantic-awareness, an early fusion module is adopted before Transformer encoder following [74], which takes image feature from backbone and text embedding as input and perform bi-directional cross-attention between them. In line with previous segmentation models [7,27,32], a pixel embedding map $M_p \in \mathcal{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ at 1/4 resolution is constructed by upsampling and integrating multi-scale feature maps from the backbone and the pixel decoder. Eventually, we dot product each object query or part query with the pixel embedding map to derive an output mask $m \in \mathcal{R}^{\frac{H}{4} \times \frac{W}{4}}$:

$$m = FFN(q_l) \otimes M_p, \quad l \in \{obj, part\} \tag{1}$$

where FFN is comprised of 3 layers feed forward network with ReLU activation functions and linear layers.

3.2 Parsing Objects into Parts

We propose a Q-Former to establish the hierarchical relationship between objects and parts. As Various common objects tend to manifest shared attributes in their constituent parts, for example, both lizards and birds exhibit similar components, such as heads and torsos. Thus, we initialize a set of query embeddings in the Q-Former to parse any object into semantic parts. We denote these universal parsing query embeddings as $q_{parse} \in \mathcal{R}^{L \times C}$, where L represents the number of the parsing queries. As shown in Fig. 2, the Q-Former is comprised of M cascaded attention modules, each module includes a self-attention layer, a cross-attention layer, and a feed forward network. The universal parsing queries are first fed into the self-attention layer and then perform cross-attention with the object queries. Note that every object query is interacted with all universal parsing queries. Hence, assume N object queries($q_{obj} \in \mathcal{R}^{N \times C}$) are generated from the object decoder, and L universal parsing queries $q_{parse} \in \mathcal{R}^{L \times C}$ are initialized in the Q-Former, we obtain $N \cdot L$ part-level queries which can be denoted as $q_{part} \in \mathcal{R}^{N \cdot L \times C}$. We refer to this process as:

$$q_{part} = Q\text{-}Former(q_{parse}; q_{obj}) \tag{2}$$

Our proposed Q-Former functions as a decomposer, extracting and representing parts from object queries. Hence, by training jointly on extensive object-level datasets and limited hierarchical datasets which contain object-part correspondences, our Q-Former obtains strong generalization ability to parse any novel object into its corresponding parts.

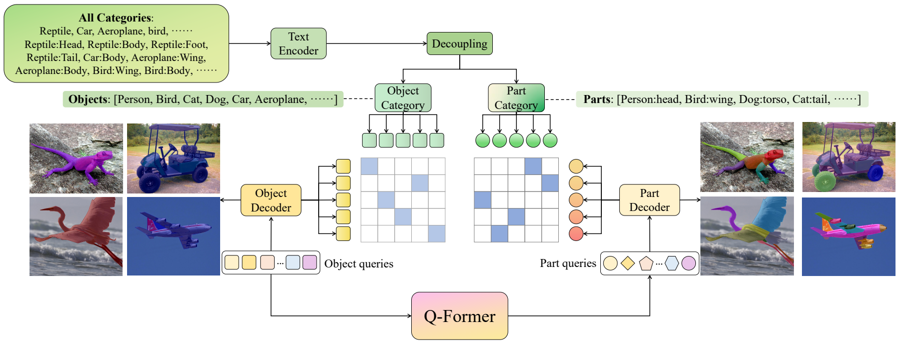
All Categories:
Reptile, Car, Aeroplane, bird, ······
Reptile:Head, Reptile:Body, Reptile:Foot,
Reptile:Tail, Car:Body, Aeroplane:Wing,
Aeroplane:Body, Bird:Wing, Bird:Body, ······
Text Encoder
Decoupling
Object Category
Part Category
Objects: [Person, Bird, Cat, Dog, Car, Aeroplane, ······]
Parts: [Person:head, Bird:wing, Dog:torso, Cat:tail, ······]
Object Decoder
Part Decoder
Object queries
Part queries
Q-Former

Fig. 3: Matching mechanisms of PartGLEE. Two separate forward passes are performed on the same image to obtain hierarchical segmentation results.

3.3 Unified Training Paradigm for Objects and Parts

Since the Q-Former requires hierarchical data to learn how to parse objects into parts, we enrich part-level data with corresponding object-level annotations, details of which are provided in the supplementary materials.

Since the annotation granularity across part-level datasets is standardized, our model can first learn the characteristics of objects and then acquire the ability to parse any object into its semantic parts. To facilitate open-vocabulary detection and segmentation, we substitute the similarity scores between the instance embeddings and the text embeddings for the original class head. Given K object-level and part-level input categories as separate sentences, we feed them into the text encoder and utilize the average of each individual sentence tokens as the output text embedding T_l for each category. Then the similarity scores $S_l \in \mathcal{R}^{N\times K}$ are calculated through a dot product operation between the hierarchical instance embeddings $q_l \in \mathcal{R}^{N\times C}$ from detector and the text embeddings $T_l \in \mathcal{R}^{K\times D}$ from text encoder, which can be denoted as:

$$S_l = q_l \cdot W_{proj} \otimes T_l, \quad l \in \{obj, part\} \tag{3}$$

where $W_{proj} \in \mathcal{R}^{C\times D}$ is a trainable projection weight for fine-tuning text embedding space especially for part-level descriptions. Following [26, 68], we perform Hungarian matching of objects and parts individually, where object-level predictions are only matched with object-level targets, and the same applies to the part-level output, as shown in Fig. 3.

We then introduce a constraint loss to ensure the part-level predictions to be the component of the objects. We denote this novel loss function as **restriction loss** L_{res}. Due to memory limitations, we only calculate our restriction loss on the predicted bounding boxes between different levels, while leaving the predicted masks unconstrained. Our restriction loss can be calculated as follow:

$$L_{res} = \sum_{i}^{L}(1 - \frac{|S_{obj} \cap S_{part}^{i}|}{S_{part}^{i}}) \tag{4}$$

where S_{obj} represents the area of the object-level bounding box prediction, and S^{i}_{part} stands for the area of the $i-th$ part-level bounding box prediction. Note that each object query can generate L part queries through Q-Former. This loss function is only applied to the matched predictions in part-level datasets, thereby strengthening the mutual correspondence between different hierarchies.

PartGLEE is trained with a linear combination of losses for object-level tasks and part-level tasks, which can be formulated as:

$$L = \lambda_1(L^{obj}_{cls} + L^{part}_{cls}) + \lambda_2(L^{obj}_{box} + L^{part}_{box}) + \lambda_3(L^{obj}_{mask} + L^{part}_{mask}) + \lambda_4 L_{res} \quad (5)$$

where L^{l}_{cls}, L^{l}_{box}, L^{l}_{mask} are the classification, box, and mask loss at different levels ($l \in \{obj, part\}$), while L_{res} is the restriction loss, and λ are their corresponding weights. We apply Focal Loss [36] as the classification loss on the similarity scores S_l to align the text concepts with instance features. A combination of L1 loss and generalized IoU loss [59] is utilized for box predictions, while we employ both Dice Loss [47] and Focal Loss to calculate mask loss. We follow MaskDINO to set our hyperparameters to $\lambda_1 = 4, \lambda_2 = 2, \lambda_3 = 5, \lambda_4 = 5$. Based on the above designs, PartGLEE is able to leverage both object-level data and part-level data thus obtaining a strong generalization capability.

4 Experiments

4.1 Experimental Setup

We conduct comprehensive experiments to exhibit the effectiveness of PartGLEE across a wide range of object-level and part-level tasks.

Data Unification. We utilize object-level datasets such as COCO [37], LVIS [16], Object365 [61], OpenImages [25], Visual Genome [24] and RefCOCO series [44,80], etc, while using part-level datasets PACO [55], PartImageNet [18], Pascal Part [6], ADE20K-Part [70] and SA-1B [23] with varying annotation granularity for joint-training. For Visual Genome and SA-1B, we categorize their corresponding part-level annotations based on semantic and mask overlap relationships to construct hierarchical data versions. For part-level data, we integrate the original part-level annotations with corresponding object-level annotations according to their associated object-level dataset. The details of these dataset preprocessing steps are left in the supplementary materials.

Implementation Details. In our experiments, we utilize ResNet-50 [20] and Swin-Large [40] as the vision encoder. Following MaskDINO [27], we adopt deformable transformer in the decoder, and use 300 object queries while setting the number of parsing queries L to be 10. The M of Q-Former is set to 6. We select the top 50 object queries based on the similarity scores and input them into the Q-Former, ultimately yielding 500 part queries. We use both query denoising and hybrid matching strategies to facilitate convergence and enhance performance. We conduct experiments on part-level datasets following the methodologies of VLPart [62] and OV-PARTS [70] in order to evaluate the generalization performance and the ability to parse novel objects of our model. For joint-training, we

Table 1: Cross-dataset generalization performance compared with VLPart. The evaluation metric is mAP_{mask} on the validation set of PartImageNet. All models utilize ResNet-50 as backbone and use the text embeddings of the category names as the classifier. PartImageNet denotes the fully-supervised method reported for comparison.

Method	Datasets	All	*quadruped*			
		(40)	*head*	*body*	*foot*	*tail*
VLPart [62]	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
PartGLEE	Pascal Part	9.9	23.6	4.5	1.3	4.6
	+ Parsed IN-S11	14.9	55.3	27.2	7.0	23.6
	vs. baseline	+5.0	+31.7	+22.7	+5.7	+19.0
	PartImageNet	40.2	67.0	37.6	36.5	40.7

train PartGLEE based on the weights of GLEE [71], continuing training on 32 A100 GPUs. The settings for the part-level zero-shot experiments are described separately in each section.

4.2 Zero-Shot Part Segmentation Results

1) Cross-dataset Part Segmentation on PartImageNet. We follow VLPart [62] to conduct experiments on cross-dataset generalization performance by directly evaluating on PartImageNet [18] validation set. We report the metrics of all (40) part categories and the detailed metrics of *quadruped* are also provided. The baseline approach only utilize Pascal Part as the training set and directly perform evaluation on PartImageNet in a zero-shot manner. Note that **IN-S11 label** represents adding image-level classification data for training in order to improve performance. Meanwhile, **Parsed IN-S11** stands for training with the pseudo-labels generated from the parsing pipeline proposed by VLPart. However, both of these methods expose the model to categories and images from the PartImageNet dataset. We first perform our training process exclusively on the Pascal Part dataset to verify our zero-shot capabilities, and then we incorporate pseudo-labels to assess the ability of our model to utilize low-quality annotations.

Given that Pascal Part does not provide semantic labels for categories like *quadruped* in PartImageNet, the model needs to generalize from annotated parts of *dog*, *cat*, etc. in Pascal Part to parts of *quadruped* in PartImageNet. As shown in Tab. 1, our model significantly outperform VLPart when only training on Pascal Part, even surpassing the model trained with **Parsed IN-S11**. After incorporating pseudo-labeled data into training, our model shows higher performance gains, indicating better utilization of low-quality data. This result illustrates the importance of hierarchical modeling, which enables our model to recognize and

Table 2: Cross-category generalization performance compared with VLPart. The evaluation metric is mAP_{mask} on the validation set of PascalPart but report $AP50$ specifically for dog parts. All models utilize ResNet-50 as backbone and use the text embeddings of the category names as the classifier. Base part represents the base split from Pascal Part. VOC object is added to the training process to improve the cognitive ability of the model thus reach a better performance. Pascal Part denotes the fully-supervised method reported for comparison.

Method	Datasets	All AP	BaseAP	NovelAP	*dog*					NovelAP
		(93)	(77)	(16)	*head*	*torso*	*leg*	*paw*	*tail*	Increment
VLPart [62]	Base Part	15.0	17.8	1.5	6.1	7.9	2.9	13.8	3.2	-
	+ VOC object	16.8	19.9	2.1	29.9	22.6	3.2	12.4	2.1	*+0.6*
	+ IN-S20 label	17.4	20.8	1.1	12.8	17.8	2.0	5.9	0.9	*-0.4*
	+ Parsed IN-S20	18.4	21.3	4.2	28.7	34.8	17.2	5.7	14.3	*+2.7*
	Pascal Part	19.4	18.8	22.4	88.0	49.6	38.3	48.9	25.8	-
PartGLEE	Base Part	25.6	30.5	2.1	12.6	15.6	8.2	5.2	6.2	-
	+ VOC object	26.9	31.2	5.8	46.5	35.0	27.0	14.7	15.1	*+3.7*
	+ Parsed IN-S20	26.6	28.9	15.5	80.3	57.3	36.7	17.0	37.4	*+9.7*
	Pascal Part	35.5	34.6	39.9	95.9	88.5	75.0	76.7	72.9	-

parse novel objects into their corresponding parts based on the generalization capability brought by CLIP.

2) Cross-category Part Segmentation on Pascal Part. We follow the evaluation setting proposed by VLPart to assess the cross-category generalization performance of our model on the Pascal Part dataset. A total of 93 part categories are divided into 77 base part categories and 16 novel part categories. Tab. 2 presents the evaluation results for all (93), base (77), and novel (16) parts. The model is trained only on the base categories, and is directly evaluated on the entire datasets. Note that **IN-S20 label** represents adding image-level classification data and **Parsed IN-S20** is on behalf of he pseudo-labels generated by VLPart [62] on ImageNet [8]. We further introduce a metric called **NovelAP Increment** on top of VLPart to assess the improvement of our model when adding extra object datasets into the training process. It is calculated by subtracting the baseline Novel AP from the Novel AP achieved after incorporating extra datasets. The results shown in Tab. 2 demonstrate that our method surpasses the performance of VLPart by a large margin. By comparing the NovelAP Increment, we observe that our method achieves a greater increment after incorporating extra object dataset. Since the VOC dataset includes object categories corresponding to novel parts, the hierarchical relationships of the Q-Former can be extended to novel part categories, resulting in a higher NovelAP Increment.

3) Generalized Zero-Shot Part Segmentation. We adopt the **Oracle-Obj setting** proposed by OV-PARTS [70] to conduct experiments on ADE-Part-234 and Pascal-Part-116 datastes. This setting assumes that the ground-truth masks and categories of object-level instances are known during the inference process, aiming to evaluate the capability of the model to parse any novel object. All categories in the datasets are divided into a base set and a novel set,

Table 3: Generalized zero-Shot part segmentation performance on ADE-Part-234 and Pascal-Part-116 compared with baselines proposed by OV-PARTS.

Method	Model	Backbone	Finetuning	Oracle-Obj					
				ADE-Part-234			Oracle-obj		
				Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
Fully Supervised	Mask2Former	ResNet-50	✕	46.25	47.86	-	55.28	52.14	-
Two-Stage	ZSseg+	ResNet-50	CPTCoOp	43.19	27.84	33.85	55.33	19.17	28.48
		ResNet-50	CPTCoCoOp	39.67	25.15	30.78	54.43	19.04	28.21
		ResNet-101c	CPTCoOp	43.41	25.70	32.28	57.88	21.93	31.81
One-Stage	CATSeg	ResNet-101&ViT-B/16	✕	11.49	8.56	9.81	14.89	10.29	12.17
		ResNet-101&ViT-B/16	B+D	31.40	25.77	28.31	43.97	26.11	32.76
	CLIPSeg	ViT-B/16	✕	15.27	18.01	16.53	22.33	19.73	20.95
		ViT-B/16	VA+L+F+D	38.96	29.65	33.67	48.68	27.37	35.04
	PartGLEE	ResNet-50	✕	51.29	35.33	41.83	57.43	27.41	37.11

and the training process is performed only on the base set, while we evaluate the performance of the model on all categories. As shown in Tab. 3, our model achieves a superior performance on both datasets, which indicates the importance of hierarchical modeling. The establishment of hierarchical relationships between objects and parts enables our model to extend to novel objects, thereby effectively parsing them into corresponding semantic parts. Consequently, our model exhibits outstanding performance across both datasets.

4.3 Joint-training results on Detection and Segmentation

To endow our model with robust generalization capability, we perform joint training on various datasets and evaluate its performance on both object-level and part-level tasks. We compare our model with specialist and generalist models to evaluate its performance on object-level data. Additionally, we contrast it with VLPart to assess its performance on part-level datasets as well as the effectiveness of joint-training process on both types of datasets. As shown in Tab. 4, PartGLEE significantly outperforms VLPart on both object-level and part-level tasks after joint-training, while achieving comparable performance on object-level tasks compared with previous SOTA. Through joint-training, our model has acquired strong generalization performance, allowing it to simultaneously address tasks for different hierarchies. We also observe that VLPart fails to achieve satisfactory performance on both object-level and part-level tasks. For example, VLPart obtains better performance on Pascal Part than its dataset-specific oracle, while decreasing its performance on COCO and LVIS. We attribute the performance drop of VLPart to the absence of hierarchical relationships, which causes confusion in modeling parts and objects and impairs object-level performance. PartGLEE effectively addresses this problem and extends the generalization capabilities from object-level to part-level tasks.

Table 4: Joint-Training Performance of PartGLEE. Note that Oracle represents the dataset-specific training paradigm. We directly evaluate the generalist models on PACO to assess their recognition capability at the part level, as indicated by the results annotated in the grey font.

Type	Method	Part-level Tasks							Object-level Tasks					
		PartImageNet		Pascal Part		PACO			COCO-val		LVIS-minival		LVIS-val	
		AP_{box}	AP_{mask}	AP_{box}	AP_{mask}	AP_{mask}	AP^{obj}_{mask}	AP^{opart}_{mask}	AP_{box}	AP_{mask}	AP_{box}	AP_{mask}	AP_{box}	AP_{mask}
Specialist	Mask2Former(R50) [7]	-	-	-	-	-	-	-	46.2	43.7	-	-	-	-
	Mask2Former(L) [7]	-	-	-	-	-	-	-	-	50.1	-	-	-	-
	MaskDINO(R50) [27]	-	-	-	-	-	-	-	50.5	46.0	-	-	-	-
	MaskDINO(L) [27]	-	-	-	-	-	-	-	58.3	52.1	-	-	-	-
	ViTDet-L [32]	-	-	-	-	-	-	-	57.6	49.8	-	-	51.2	46.0
	ViTDet-H [32]	-	-	-	-	-	-	-	57.6	49.8	-	-	53.4	48.1
	EVA-02-L [12]	-	-	-	-	-	-	-	64.2	55.0	-	-	65.2	57.3
	PACO(R50) [55]	-	-	-	-	-	32.6	12.5	-	-	-	-	-	-
	PACO(L) [55]	-	-	-	-	-	43.4	17.7	-	-	-	-	-	-
Generalist	Pix2Seq v2 [5]	-	-	-	-	-	-	-	46.5	38.2	-	-	-	-
	X-Decoder(L) [89]	-	-	-	-	2.69	11.9	0.94	-	46.7	-	-	-	-
	SEEM(L) [90]	-	-	-	-	1.99	8.42	0.69	-	47.7	-	-	-	-
	HIPIE(R50) [68]	-	-	-	-	-	-	-	53.9	45.9	-	-	-	-
	Florence-2(B) [72]	-	-	-	-	-	-	-	41.4	-	-	-	-	-
	Florence-2(L) [72]	-	-	-	-	-	-	-	43.4	-	-	-	-	-
	UNINEXT(R50) [74]	-	-	-	-	-	-	-	51.3	44.9	-	-	36.4	-
	UNINEXT(L) [74]	-	-	-	-	-	-	-	58.1	49.6	-	-	-	-
	GLEE(R50) [71]	-	-	-	-	3.44	15.3	1.29	55.0	48.4	50.5	45.9	44.2	40.2
Hierarchical	VLPart(R50) [62]	30.7	31.6	23.9	24.0	13.8	36.9	9.6	28.5	-	-	26.2	-	-
	VLPart(R50)-Oracle [62]	29.2	29.7	18.9	19.4	13.3	28.0	10.6	38.0	-	-	28.1	-	-
	VLPart(B) [62]	43.9	41.2	33.5	31.7	22.1	55.0	15.9	40.3	-	-	39.6	-	-
	VLPart(B)-Oracle [62]	44.3	41.7	29.2	27.4	19.1	37.7	15.2	52.5	-	-	43.1	-	-
	PartGLEE (R50)	40.9	40.2	35.0	35.5	21.8	50.5	15.4	54.4	47.6	48.7	43.5	42.7	38.3
	PartGLEE (L)	52.7	50.9	39.6	39.1	27.8	55.7	21.3	59.5	52.0	56.5	50.6	50.2	45.0

4.4 Ablation Study

To demonstrate that our model design achieves satisfactory results on both object-level and part-level tasks, we conduct an ablation study (depicted in Fig. 4) on the model architecture and present results in Tab. 5. We ablate with a backbone of ResNet-50 and perform joint-training on COCO [37], LVIS [16], PartImageNet [18], Pascal Part [6] and PACO [55] with 90K iterations. From this study, we draw several important conclusions: (1) The utilization of parallel pixel decoders only results in slight improvements in mask predictions on few datasets, indicating that the influence of feature maps at different granularities is negligible. (2) Adopting independent decoders to obtain predictions at different levels demonstrates superior performance across the majority of datasets, manifesting the effectiveness of independent decoders. As adopting parallel pixel decoders (b) results in significant GPU memory costs without considerable gains, and all metrics for (a) are lower than (c), we select (c) as our final model design. Additional ablation studies, extensive qualitative analysis, and experiments on mLLM can be found in the supplementary materials.

4.5 Limitations

In this work, we still adopt CLIP as the text encoder, which is trained on text-image pairs and thus lacks the ability to perceive fine-grained descriptions of object or part instances. This limitation may restrict the improvement of model

Table 5: An ablation study on different model designs, as depicted in Fig 4. Note that Parallel Pixel Decoders refers to the utilization of two pixel decoders to generate feature maps at different hierarchies respectively. Independent Decoders denote the usage of two decoders, which facilitate the interaction between feature maps and queries at different hierarchies. Our final choice is scheme (c), which is highlighted in gray.

Scheme	Model Design		Part-level Tasks					Object-level Tasks			
	Parrallel Pixel Decoders	Independent Decoders	PartImageNet	Pascal Part	PACO			COCO-val		LVIS-minival	
			AP_{mask}	AP_{mask}	AP_{mask}	AP^{obj}_{mask}	AP^{opart}_{mask}	AP_{box}	AP_{mask}	AP_{box}	AP_{mask}
(a)	✗	✗	**39.0**	34.1	20.1	47.4	13.5	47.8	43.5	34.8	33.4
(b)	✓	✗	38.3	34.5	20.8	**48.8**	13.8	48.5	**44.3**	34.9	**34.2**
(c)	✗	✓	**39.0**	**34.7**	**20.9**	47.9	**14.2**	**49.3**	44.2	**35.6**	33.8

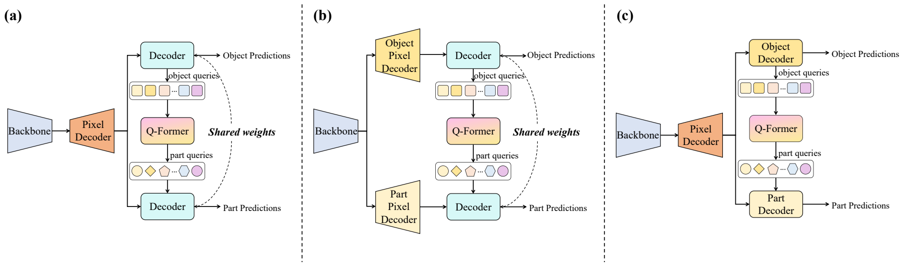

Fig. 4: Various designs for generating predictions at different hierarchies. In scheme (a), we only utilize a single decoder to generate predictions for both objects and parts. In scheme (b), two parallel pixel decoders are employed to generate feature maps at different levels, aiming to explore the effectiveness of feature maps at different granularity. In scheme (c), we use two independent decoders to generate predictions for objects and parts respectively.

performance and prompts us to consider how to enhance the perception capabilities of region-level models, which will be our future work.

5 Conclusion

In this paper, we introduce PartGLEE, a groundbreaking foundation model designed towards a complete comprehension of both objects and parts in images. Through the generic hierarchical relationships established by the Q-Former, we are able to break through the limitation of scarce part-level data by introducing a large amount of object-level data, thereby transferring the powerful generalization capabilities from objects to parts. Through extensive training on diverse datasets, PartGLEE achieves SOTA performance across various part-level tasks while maintaining competitive results on object-level tasks, enabling it to parse any objects into parts and serve as a foundation model for general fine-grained region-level perception tasks.

Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No.62225603).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Advances in neural information processing systems. vol. 35, pp. 23716–23736 (2022)
2. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in neural information processing systems. vol. 33, pp. 1877–1901 (2020)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chen, T., Saxena, S., Li, L., Lin, T.Y., Fleet, D.J., Hinton, G.E.: A unified sequence interface for vision tasks. Advances in Neural Information Processing Systems **35**, 31333–31346 (2022)
6. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: 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. pp. 1971–1978 (2014)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 248–255 (2009)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S.: Towards unified human parsing and pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 843–850 (2014)
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
12. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023)
13. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)

14. de Geus, D., Meletis, P., Lu, C., Wen, X., Dubbelman, G.: Part-aware panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5485–5494 (2021)
15. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: 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. pp. 932–940 (2017)
16. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019)
17. He, J., Chen, J., Lin, M.X., Yu, Q., Yuille, A.L.: Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11259–11268 (2023)
18. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: European Conference on Computer Vision. pp. 128–145. Springer (2022)
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
25. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
26. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
27. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023)

28. Li, H., Zhu, J., Jiang, X., Zhu, X., Li, H., Yuan, C., Wang, X., Qiao, Y., Wang, X., Wang, W., et al.: Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2691–2700 (2023)
29. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J.: Multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206 (2017)
30. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
31. Li, X., Xu, S., Yang, Y., Cheng, G., Tong, Y., Tao, D.: Panoptic-partformer: Learning a unified model for panoptic part segmentation. In: European Conference on Computer Vision. pp. 729–747. Springer (2022)
32. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022)
33. Li, Y., Singh, K.K., Xue, Y., Lee, Y.J.: Partgan: Weakly-supervised part decomposition for image generation and segmentation. In: British Machine Vision Conference (BMVC) (2021)
34. Lin, C., Jiang, Y., Qu, L., Yuan, Z., Cai, J.: Generative region-language pretraining for open-ended object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13958–13968 (2024)
35. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: The Eleventh International Conference on Learning Representations (2023)
36. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
38. Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: Editgan: High-precision semantic image editing. In: Advances in Neural Information Processing Systems. vol. 34, pp. 16331–16345 (2021)
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
41. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. In: The Eleventh International Conference on Learning Representations (2022)
42. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In: Advances in neural information processing systems. vol. 36 (2023)
43. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. arXiv preprint arXiv:2404.13013 (2024)

44. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
45. Meletis, P., Wen, X., Lu, C., de Geus, D., Dubbelman, G.: Cityscapes-panoptic-parts and pascal-panoptic-parts datasets for scene understanding. arXiv preprint arXiv:2004.07944 (2020)
46. Michieli, U., Borsato, E., Rossi, L., Zanuttigh, P.: Gmnet: Graph matching network for large scale part semantic segmentation in the wild. In: European Conference on Computer Vision. pp. 397–414. Springer (2020)
47. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV) (2016)
48. Morabia, K., Arora, J., Vijaykumar, T.: Attention-based joint detection of object and semantic part. arXiv preprint arXiv:2007.02419 (2020)
49. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. In: Conference on Robot Learning. pp. 892–909. PMLR (2023)
50. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19023–19034 (2022)
51. Pan, T.Y., Liu, Q., Chao, W.L., Price, B.: Towards open-world segmentation of parts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15392–15401 (2023)
52. Qi, L., Kuen, J., Guo, W., Gu, J., Lin, Z., Du, B., Xu, Y., Yang, M.H.: Aims: All-inclusive multi-level segmentation for anything. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
53. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
54. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
55. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7151 (2023)
56. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
57. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
58. Reddy, N.D., Vo, M., Narasimhan, S.G.: 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. pp. 1906–1915 (2018)
59. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)

60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
61. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
62. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15453–15465 (2023)
63. Tang, C., Xie, L., Zhang, X., Hu, X., Tian, Q.: Visual recognition by request. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15265–15274 (2023)
64. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds-200-2011 (cub-200-2011). Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
65. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Joint object and part segmentation using deep learned potentials. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1573–1581 (2015)
66. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
67. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)
68. Wang, X., Li, S., Kallidromitis, K., Kato, Y., Kozuka, K., Darrell, T.: Hierarchical open-vocabulary universal image segmentation. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
69. Wang, Y., Xu, Y., Tsogkas, S., Bai, X., Dickinson, S., Siddiqi, K.: Deepflux for skeletons in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5287–5296 (2019)
70. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: Ov-parts: Towards open-vocabulary part segmentation. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
71. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: General object foundation model for images and videos at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3783–3795 (2024)
72. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4818–4829 (2024)
73. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: European Conference on Computer Vision (2022)
74. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15325–15336 (2023)
75. Yang, L., Song, Q., Wang, Z., Jiang, M.: Parsing r-cnn for instance-level human analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 364–373 (2019)

76. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1385–1392. IEEE (2011)
77. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision. pp. 521–539. Springer (2022)
78. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023)
79. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In: Advances in Neural Information Processing Systems. vol. 35, pp. 9125–9138 (2022)
80. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85. Springer (2016)
81. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
82. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
83. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
84. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal on Computer Vision (2018)
85. Zhou, T., Wang, W., Liu, S., Yang, Y., Van Gool, L.: 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. pp. 1622–1631 (2021)
86. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)
87. Zhu, X., Zhu, J., Li, H., Wu, X., Li, H., Wang, X., Dai, J.: Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16804–16815 (2022)
88. Ziegler, A., Asano, Y.M.: Self-supervised learning of object parts for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14502–14511 (2022)
89. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)
90. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: Advances in Neural Information Processing Systems. vol. 36 (2023)