General Object Foundation Model for Images and Videos at Scale

Junfeng Wu[1*], Yi Jiang[2*], Qihao Liu[3], Zehuan Yuan[2], Xiang Bai[1†], Song Bai[2†]
[1]Huazhong University of Science and Technology, [2]ByteDance Inc., [3]Johns Hopkins University

Abstract

We present GLEE in this work, an object-level foundation model for locating and identifying objects in images and videos. Through a unified framework, GLEE accomplishes detection, segmentation, tracking, grounding, and identification of arbitrary objects in the open world scenario for various object perception tasks. Adopting a cohesive learning strategy, GLEE acquires knowledge from diverse data sources with varying supervision levels to formulate general object representations, excelling in zero-shot transfer to new data and tasks. Specifically, we employ an image encoder, text encoder, and visual prompter to handle multi-modal inputs, enabling to simultaneously solve various object-centric downstream tasks while maintaining state-of-the-art performance. Demonstrated through extensive training on over five million images from diverse benchmarks, GLEE exhibits remarkable versatility and improved generalization performance, efficiently tackling downstream tasks without the need for task-specific adaptation. By integrating large volumes of automatically labeled data, we further enhance its zero-shot generalization capabilities. Additionally, GLEE is capable of being integrated into Large Language Models, serving as a foundational model to provide universal object-level information for multi-modal tasks. We hope that the versatility and universality of our method will mark a significant step in the development of efficient visual foundation models for AGI systems. The models and code are released at https://github.com/FoundationVision/GLEE.

1. Introduction

Foundation model [7] is an emerging paradigm for building artificial general intelligence (AGI) systems, signifying a model trained on broad data that is capable of being adapted to a wide range of downstream tasks in an general paradigm. Recently, NLP foundation models such as BERT [21], GPT-3 [9], T5 [70] developed with unified

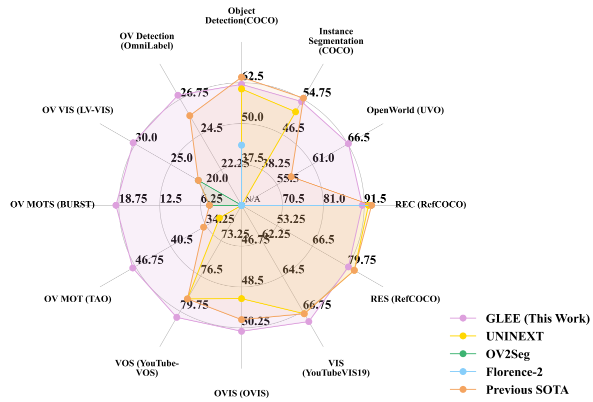

Figure 1. The performance of GLEE on a broad range of object-level tasks compared with existing models.

input-output paradigms and large-scale pre-training, have achieved remarkable generalization capabilities to address nearly all NLP tasks.

In computer vision, the diversity of task types and the lack of a unified form make visual foundation models only serve specific subdomains, such as CLIP [69] for multi-modal visual model, MAE [32] for visual representations model, SAM [39] for segmentation model. Despite being widely studied, current visual foundation models are still focusing on establishing correlations between global image features and language descriptions or learning image-level feature representations. However, locating and identifying objects constitute foundational capabilities in computer vision systems, serves as a basis for solving complex or high level vision tasks such as segmentation, scene understanding, object tracking, event detection, and activity recognition and support a wide range of applications.

In this work, we advance the development of object-level foundation models within the visual domain. To address the aforementioned limitation, providing general and accurate object-level information, we introduce a general object visual foundation model, coined as GLEE, which simultaneously solve a wide range of object-centric tasks while ensuring SOTA performance, including object detection, instance segmentation, grounding, object tracking, interactive segmentation and tracking, etc., as shown in Figure 1. Through

[*]Equal Contribution. This work was done when Junfeng Wu and Qihao Liu were interns at ByteDance. [†]Correspondence to Xiang Bai <xbai@hust.edu.cn> and Song Bai <songbai.site@gmail.com>.

a unified input and output paradigm definition, our model is capable of learning from a wide range of diverse data and predicting general object representations, which makes it generalize well to new data and tasks in a zero-shot manner and achieve amazing performance. In addition, thanks to the unified paradigm, the training data can be scaled up at low cost by introducing a large amount of automatically labeled data, and further improve the zero-shot generalization ability of the model.

A general object foundation model framework. Our objective is to build an object visual foundation model capable of simultaneously addressing a wide range of object-centric tasks. Specifically, we employ an image encoder, a text encoder, and a visual prompter to encode multi-modal inputs. They are integrated into a detector to extract objects from images according to textual and visual input. This unified approach to handle multiple modalities enables us to concurrently solve various object-centric tasks, including detection [10, 52, 81, 119], instance segmentation [15, 31], referring expression comprehension [35, 55, 93, 118], interactive segmentation [1, 12, 122], multi-object tracking [20, 60, 99, 113, 116], video object segmentation [16, 17, 65, 98], video instance segmentation [34, 87, 90, 92, 102], and video referring segmentation [77, 91, 93], all while maintaining state-of-the-art performance.

Multi-granularity joint supervision and scalable training paradigm. The design of the unified framework capable of addressing multiple tasks enables joint training on over five million images from diverse benchmarks and varying levels of supervision. Existing datasets differ in annotation granularity: detection datasets like Objects365 [79] and OpenImages [42] offer bounding boxes and category names; COCO [52] and LVIS [29] provide finer-grained mask annotations; RefCOCO [64, 108] and Visual Genome [40] provide detailed object descriptions. Additionally, video data enhance the temporal consistency of model, while open-world data contribute class-agnostic object annotations. An intuitive display of the supervision types and data scales of the datasets employed is presented in Figure 2. The unified support for multi-source data in our approach greatly facilitates the incorporation of additional manually or automatically annotated data, enabling easy scaling up of the dataset. Furthermore, the alignment of model optimization across tasks means that joint training serves not only as a unifying strategy but also as a mechanism to boost performance across individual tasks.

Strong zero-shot transferability to a wide range of object level image and video tasks. After joint training on data from diverse sources, GLEE demonstrates remarkable versatility and zero-shot generalization abilities. Extensive experiments demonstrate that GLEE achieves state-of-the-art performance compared to existing specialist and generalist models in object-level image tasks such as detection, referring expression comprehension, and open-world detection, all without requiring any task-specific designs or fine-tuning. Furthermore, we showcase the extraordinary generalization and zero-shot capabilities of GLEE in large-vocabulary open-world video tracking tasks, achieving significantly superior performance over existing models even in a zero-shot transfer manner. Additionally, by incorporating automatically annotated data like SA1B [39] and GRIT [67], we are able to scale up our training dataset to an impressive size of 10 million images at a low cost, which is typically challenging to achieve for object-level tasks and further enhances the generalization performance. Moreover, we replace the SAM [39] component with GLEE in a multimodal Large Language Model (mLLM) [43] and observe that it achieves comparable results. This demonstrates that GLEE is capable of supplying the visual object-level information that modern LLMs currently lack, thus laying a solid foundation for an object-centric mLLMs.

2. Related Work

2.1. Visual Foundation Model

As foundation models [9, 18, 21, 70, 82] in the NLP field have achieved remarkable success, the construction of visual foundation models attracts increasing attention. Unlike NLP tasks that are predominantly unified under a text-to-text paradigm, tasks in Computer Vision still exhibit significant differences in formulation. This disparity leads to the fact that visual models are typically trained in a single-task learning frameworks, limiting their applicability to tasks within certain sub-domains. For instance, multi-modal visual foundation models like CLIP [69], ALIGN [37], Florence [109], BEIT3 [86], Flamingo[2] make significant advancements in efficient transfer learning and demonstrate impressive zero-shot capabilities on vision-language tasks by employing contrastive learning and masked data modeling on large-scale image-text pairs. DALL-E [71, 72] and Stable Diffusion [74] are trained on massive pairs of images and captions, enabling them to generate detailed image content conditioned on textual instruction. DINO [11], MAE [32], EVA [25], ImageGPT [13] obtain strong visual representations through self-supervised training on large-scale image data, which are then employed to adopt downstream tasks. These foundation models learn image-level features and are not directly applicable to object-level tasks. The recently proposed SAM [39], capable of segmenting any object of a given image based on visual prompt such as points and boxes, provides rich object-level information and demonstrates strong generalization capabilities. However, the object information lacks semantic context, limiting its application in object-level tasks. Unlike existing visual foundation models, we aim to develop an object foundation model that directly solve downstream tasks without the need

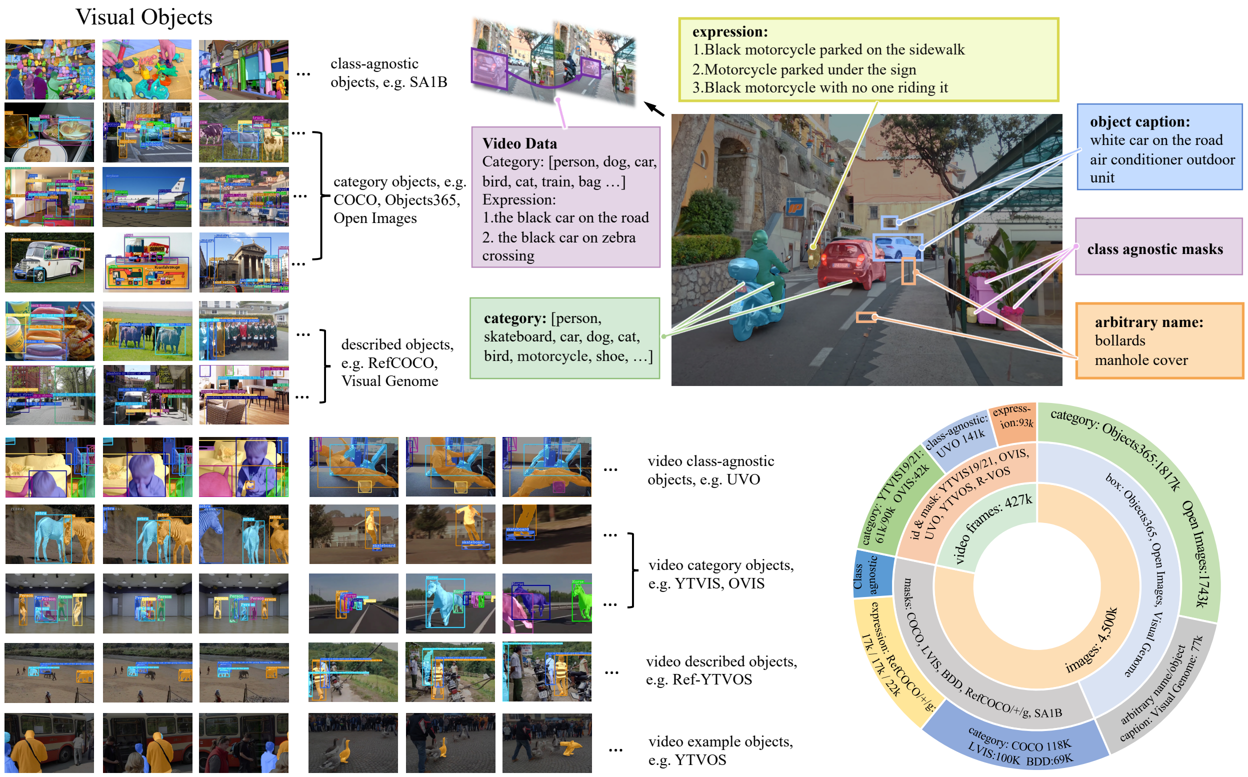

Figure 2. An illustrative example showcasing annotations of varying granularities from different datasets, along with the scale of data we utilized. Training on datasets from multiple sources endows the model with more universal representations.

for additional parameters or fine-tuning.

2.2. Unified and General Model

Unified models share similarities with foundation models in the aspect of multi-task unification for their ability to handle multiple vision or multi-modal tasks within a single model. MuST [27] and Intern [78] propose to train across multiple vision tasks and solving them simultaneously. Inspired by the success of sequence-to-sequence NLP models [9, 70], models such as Uni-Perceiver [120], OFA [84], Unified-IO [58], Pix2Seq v2 [14], and UniTAB [103] propose modeling various tasks as sequence generation tasks within a unified paradigm. While these approaches have demonstrated promising cross-task generalization capabilities, they focus mainly on image-level understanding tasks. In addition, their auto-regressive generation of boxes and masks results in significantly slower inference speeds and the performance still falls short of state-of-the-art task-specific models. Building upon on detectors [45, 119], Uni-Perceiver v2 [46] and UNINEXT [100] utilize unified maximum likelihood estimation and object retrieval to support various tasks, effectively resolves the challenges of localization. Nonetheless, they are trained on closed-set data, thereby not exhibiting zero-shot generalization capabilities. X-decoder [121] and SEEM [122] construct a generalized decoding model capable of predicting pixel-level segmentation and language tokens. Diverging from unified models, the proposed GLEE not only directly addresses object-level tasks in a unified manner but also provides universal object representations, which generalize well to new data and tasks, serving as a cornerstone for a broader range of tasks that require detailed object information.

2.3. Vision-Language Understanding

Open-vocabulary detection (OVD) and Grounding models both necessitate the localization and recognition of as many objects as possible. With the recent advancements in vision-language pre-training [37, 69, 107, 109], a commonly employed strategy for OVD involves transferring the knowledge from pre-trained vision-language models (VLMs) to object detectors [28, 41, 63]. Another group of studies leverages extensive image-text pair datasets to broaden the detection vocabulary [26, 47, 51, 59, 101, 105, 110, 115]. However, these language-based detectors are inherently constrained by the capabilities and biases of language models, making it challenging to excel simultaneously in both localization and recognition. Our objective is to optimally utilize existing datasets to construct a general object-level foundation model, aims to not only detect and identify objects effectively but also to offer universal object representations for a wide range of downstream tasks.

3. Method

3.1. Formulation

The proposed GLEE consists of an image encoder, a text encoder, a visual prompter, and an object decoder, as illustrated in Figure 3. The text encoder processes arbitrary descriptions related to the task, including object categories, names in any form, captions about objects, and referring expressions. The visual prompter encodes user inputs such as points, bounding boxes, or scribbles during interactive segmentation into corresponding visual representations of target objects. Then they are integrated into a detector for extracting objects from images according to textual and visual input. We build the object decoder upon MaskDINO [45] with a dynamic class head by compute similarity between object embedding from detector and text embedding from the text encoder. Given an input image $I \in \mathcal{R}^{3\times H\times W}$, we first extract multi-scale features Z with backbones such as ResNet [30]. Then we feed them into the object decoder and adopt three prediction heads (classification, detection, and segmentation) on the output embedding $q_d \in \mathcal{R}^{N\times C}$ from decoder. Following other object segmentation models [15, 45, 50], we construct a 1/4 resolution pixel embedding map $M_p \in \mathcal{R}^{C\times \frac{H}{4}\times\frac{W}{4}}$ which is obtained by upsampling and fusing multi-scale feature maps from the backbone and Transformer encoder. Finally, we obtain each binary mask prediction $m \in \mathcal{R}^{N\times \frac{H}{4}\times\frac{W}{4}}$ via a dot product between the N mask embeddings and pixel embedding map:

$$m = FFN(q_d) \otimes M_p, \tag{1}$$

where FFN is a 3-layer feed forward head with ReLU activation function and a linear projection layer.

To support arbitrary vocabularies and object descriptions, we replace the FFN classifier with text embeddings following DetCLIP [104]. Specifically, we feed K category names as separate sentences into the text encoder Enc_L and use the average of each sentence tokens as the output text embedding $e_t \in \mathcal{R}^{K\times D}$ for each category or description. Then we compute the alignment scores $S_{align} \in \mathcal{R}^{N\times K}$ between object embedding and text embedding:

$$S_{align} = q_d \cdot W_{i2t} \otimes e_t, \tag{2}$$

where $W_{i2t} \in \mathcal{R}^{C\times D}$ is image-to-text projection weights. We use logits S_{align} to replace traditional classification logits to compute Hungarian matching cost during training and assign categories to objects during inference. To make the original visual features prompt-aware, an early fusion module is adopted before Transformer encoder following UNINEXT [100], which takes image feature from backbone and prompt embedding as input and perform bi-directional cross-attention between them.

3.2. Task Unification

Based on the above designs, GLEE can be used to seamlessly unify a wide range of object perception tasks in images and videos, including object detection, instance segmentation, grounding, multi-target tracking (MOT), video instance segmentation (VIS), video object segmentation (VOS), interactive segmentation and tracking, and supports open-world/large-vocabulary image and video detection and segmentation tasks.

Detection and Instance Segmentation. For detection task, a fixed-length category list is given and all objects in the category list are required to be detected. For a dataset with category list length K, the text input can be formulated as $\{p_k\}_{k=1}^{K}$ where p_k represents for the k-th category name, e.g., P = ["person", "bicycle", "car", ... , "toothbrush"] for COCO [52]. For datasets with large vocabulary, calculating the text embedding of all categories is very time-consuming and redundant. Therefore, for datasets with a category number greater than 100, such as objects365 [79] and LVIS [29], suppose there are $\hat{K}$ positive categories in an image, we take the $\hat{K}$ positive categories and then pad the category number to 100 by randomly sampling from the negative categories. For instance segmentation, we enable the mask branch and add mask matching cost with mask loss.

Grounding and Referring Segmentation. These tasks provide reference textual expressions, where objects are described with attributes, for example,Referring Expression Comprehension (REC) [108, 118], Referring Expression Segmentation (RES) [55, 108], and Referring Video Object Segmentation (R-VOS) [77, 91] aim at finding objects matched with the given language expressions like "The fourth person from the left". For each image, we take the all the object expressions as text prompt and feed the them into the text encoder. For each expressions, we apply global average pooling along the sequence dimension to get text embedding e_t. The text embeddings are feed into early fusion module and additionally interacte with object queries through self-attention module in the decoder.

MOT and VIS. Both Multi-object Tracking (MOT)[4, 20, 60, 113, 116] and Video Instance Segmentation (VIS)[34, 68, 92, 102] need to detect and track all the objects in the predefined category list, and VIS requires additional mask for the objects. These two tasks can be considered as extended tasks of detection and instance segmentation on videos. We found that with sufficient image exposure, object embeddings from the decoder effectively distinguish objects in a video, showing strong discriminability and temporal consistency. As a result, they can be directly employed for tracking without the need for an additional tracking head. Training on image-level data can address

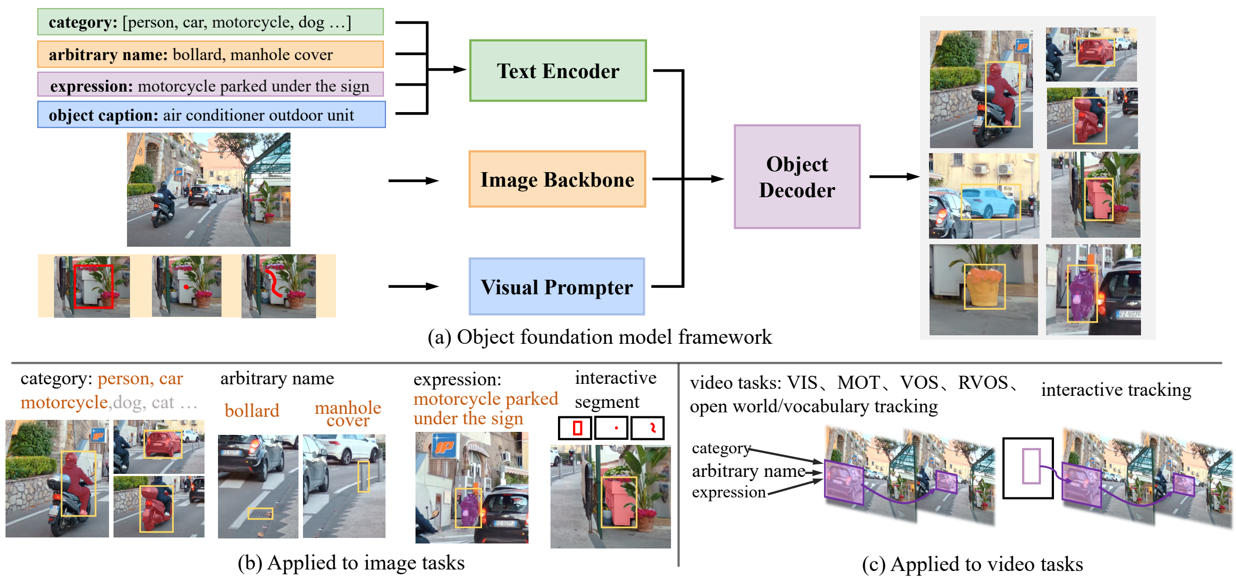

Figure 3. **Framework of GLEE.** The text encoder accepts textual descriptions in various forms from diverse data sources, including object categories, names, captions, and referring expressions. The visual prompter encodes points, bounding boxes, or scribbles into corresponding visual representations.The object decoder take them and image features to predict objects in images. (b) illustrates the application of GLEE to image tasks tailored for different language descriptions and visual prompts. (c) demonstrates the application across various object-level video tasks.

straightforward tracking scenarios, but in cases of severe occlusion scenes, such as OVIS [68], image-level training cannot guarantee that the model exhibits strong temporal consistency under occlusion conditions. Therefore, for occlusion scenarios, it is essential to utilize video data for training. Following IDOL [92], we sample two frames from a video and introduce contrastive learning between frames to make the embedding of the same object instance closer in the embedding space, and the embedding of different object instances farther away. During Inference, the detected objects are tracked by simple bipartite matching of the corresponding object queries following MinVIS [36].

Visual Prompted Segmentation. Interactive segmentation[8, 12, 56, 75, 80, 89, 97] takes various forms of visual prompt, such as points, boxes, or scribbles, to segment the specified objects within an image. On the other hand, VOS [22, 98] aims to segment the entire object throughout the entire video based on a mask provided in the first frame of the video. We extract visual prompt embeddings twice in the model. First, we crop the prompt square area from RGB image and send it to the backbone to obtain the visual prompt feature of the corresponding area, and send it to the early fusion module before the Transformer encoder. Second, we sample fine-grained visual embeddings from the pixel embedding map M_p according to visual prompt and make them interacte with object queries through self-attention module in the Transformer decoder layer, as the same with text embeddings.

3.3. Training Unification

Tasks with Dynamic Loss. We jointly train GLEE in an end-to-end manner on over 5 million images from diverse benchmarks with various levels of supervision. Different loss functions are selected for training on various datasets. There are six types of losses for our GLEE: semantic loss, box loss, mask loss, confidence loss, contrastive tracking loss, and distillation loss. For all tasks with category list or object expressions, we apply Focal loss [53] as semantic loss on logits S_{align} to align the text concepts with object features. For box prediction, we use a combination of L1 loss and generalized IoU loss [73]. The mask loss is defined as a combination of the Dice loss [62] and Focal loss. For the Visual Prompt Segmentation tasks, we employ an additional FFN to predict the confidence score for each object queries supervised by Focal loss. For video tasks fine-tuning, we sample two frames and apply contrastive tracking loss on the object query from the last layer of decoder following IDOL [92]. For the text encoder, we distill the knowledge from the frozen teacher CLIP text encoder to ensure the text embedding in pre-trained vison-language embedding space. We apply an L1 loss between our text encoder and CLIP text encoder to minimize their distance:

$$\mathcal{L}_{text} = \frac{1}{K}\sum_{i=0}^{K} \left\| Enc_L(p_k) - Enc_{CLIP}(p_k) \right\|, \quad (3)$$

where $\{p_k\}$ is the category name from the category list P.

Method	Type	*Generic Detection & Segmentation*								*Referring Detection & Segmentation*						*OpenWorld*
		COCO-val		COCO-test-dev		LVIS				RefCOCO		RefCOCO+		RefCOCOg		UVO
		AP_{box}	AP_{mask}	AP_{box}	AP_{mask}	AP_{box}	AP_{r-box}	AP_{mask}	AP_{r-mask}	P@0.5	oIoU	P@0.5	oIoU	P@0.5	oIoU	AR_{mask}
MDETR [38]		-	-	-	-	-	-	-	-	87.5	-	81.1	-	83.4	-	-
SeqTR [118]		-	-	-	-	-	-	-	-	87.0	71.7	78.7	63.0	82.7	64.7	-
PolyFormer (L) [55]		-	-	-	-	-	-	-	-	90.4	76.9	85.0	72.2	85.8	71.2	-
ViTDet-L [50]		57.6	49.8	-	-	51.2	-	46.0	34.3	-	-	-	-	-	-	-
ViTDet-H [50]	Specialist Models	58.7	50.9	-	-	53.4	-	48.1	36.9	-	-	-	-	-	-	-
EVA-02-L [24]		64.2	55.0	64.5	55.8	65.2	-	57.3	-							
ODISE [95]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	57.7
Mask2Former (L) [15]		-	50.1	-	50.5	-	-	-	-	-	-	-	-	-	-	-
MaskDINO (L) [45]		-	54.5	-	54.7	-	-	-	-	-	-	-	-	-	-	-
UniTAB (B) [103]		-	-	-	-	-	-	-	-	88.6	-	81.0	-	84.6	-	-
OFA (L) [84]		-	-	-	-	-	-	-	-	90.1	-	85.8	-	85.9	-	-
Pix2Seq v2 [14]		46.5	38.2	-	-	-	-	-	-	-	-	-	-	-	-	-
Uni-Perceiver-v2 (B) [46]		58.6	50.6	-	-	-	-	-	-	-	-	-	-	-	-	-
Uni-Perceiver-v2 (L) [46]		61.9	53.6	-	-	-	-	-	-	-	-	-	-	-	-	-
UNINEXT (R50) [100]		51.3	44.9	-	-	36.4	-	-	-	89.7	77.9	79.8	66.2	84.0	70.0	-
UNINEXT (L) [100]	Generalist Models	58.1	49.6	-	-	-	-	-	-	91.4	80.3	83.1	70.0	86.9	73.4	-
UNINEXT (H) [100]		60.6	51.8	-	-	-	-	-	-	92.6	82.2	85.2	72.5	88.7	74.7	-
GLIPv2 (B) [111]		-	-	58.8	45.8	-	-	-	-	-	-	-	-	-	-	-
GLIPv2 (H) [111]		-	-	60.6	48.9	-	-	-	-	-	-	-	-	-	-	-
X-Decoder (B) [121]		-	45.8	-	45.8	-	-	-	-	-	-	-	-	-	-	-
X-Decoder (L) [121]		-	46.7	-	47.1	-	-	-	-	-	-	-	-	-	-	-
Florence-2 (L) [94]		43.4	-	-	-	-	-	-	-	93.4	-	88.3	-	91.2	-	-
GLEE-Lite		55.0	48.4	54.7	48.3	44.2	36.7	40.2	33.7	88.5	77.4	78.3	64.8	82.9	68.8	66.6
GLEE-Plus	Foundation Models	60.4	53.0	60.6	53.3	52.7	44.5	47.4	40.4	90.6	79.5	81.6	68.3	85.0	70.6	70.6
GLEE-Pro		62.0	54.2	62.3	54.5	55.7	49.2	49.9	44.3	91.0	80.0	82.6	69.6	86.4	72.9	72.6

Table 1. Comparison of GLEE to recent specialist and generalist models on object-level image tasks. For REC and RES tasks, we report Precision@0.5 and overall IoU (oIoU). For open-world instance segmentation task, we reported the average recall of 100 mask proposals (AR@100) on the UVO [85].

Data Scale Up. A visual foundation model should be able to easily scale up the training data and achieve better generalization performance. Thanks to the unified training paradigm, the training data can be scaled up at low cost by introducing a large amount of automatically labeled data from SA1B [39] and GRIT [67]. SA1B provides large and detailed mask annotations, which enhance the general object perception capabilities of model, while GRIT offers a more extensive collection of referring-expression-bounding-box pairs, improving the object identification abilities and the understanding capability of object descriptions. Ultimately, we introduced 2 million SA1B data and 5 million GRIT data into the training process, bringing the total training data to 10 million.

4. Experiments

4.1. Experimental Setup

We conducted training in three stages. Initially, we performed pretraining for object detection task on Objects365 [79] and OpenImages [42], initializing the text encoder with pretrained CLIP [69] weights and keeping the parameters frozen. In the second training stage, we introduced additional instance segmentation datasets, including COCO [52], LVIS [29], and BDD [106]. Furthermore, we treat three VIS datasets: YTVIS19 [102], YTVIS21 [96], and OVIS [68], as independent image data to enrich the scenes. For datasets that provide descriptions of objects, we included RefCOCO [108], RefCOCO+ [108], RefCOCOg [64], Visual Genome [40], and RVOS [77]. Additionally, we introduced two open-world instance segmentation datasets, UVO [85] and a subset of SA1B [39]. Building upon this, we perform the third training stage by introducing more SA1B data and GRIT [67] data to scale up the training set, resulting in a model named **GLEE-scale**, which exhibited even stronger zero-shot performance on various downstream tasks. During the second and third stages, text encoder is unfrozen but supervised by distillation loss to ensure the predicted text embedding in CLIP embedding space. After the second step, GLEE demonstrated state-of-the-art performance on a range of downstream image and video tasks and exhibited strong zero-shot generalization capabilities, unless otherwise specified, all the experimental results presented below were obtained by the model from this stage. We developed GLEE-Lite, GLEE-Plus, and GLEE-Pro using ResNet-50 [30], Swin-Large [57], and EVA-02 Large [24] as the vision encoder respectively, and train GLEE on 64 A100 GPUs for 500,000 iterations in each stage. More detailed information on data usage, data sampling strategies, and model training can be found in the supplementary materials.

4.2. Comparison with Generalist Models

We demonstrate the universality and effectiveness of our model as an object-level visual foundation model, directly applicable to various object-centric tasks while ensuring SOTA performance without needing fine-tuning. We report detection and instance segmentation results on both the COCO-2017 [52] and LVIS val v1.0 [29]. While sharing almost identical image sets, LVIS is distinguished by

Method	Tracking Any Object (TAO [19])				BURST [3]						LV-VIS [83]		
	TETA	LocA	AssocA	ClsA	ALL		Common		Uncommon		AP	AP_b	AP_n
					HOTA	mAP	HOTA	mAP	HOTA	mAP			
Tracktor [5]	24.2	47.4	13.0	12.1	-	-	-	-	-	-	-	-	-
DeepSORT [88]	26.0	48.4	17.5	12.1	-	-	-	-	-	-	-	-	-
Tracktor++ [19]	28.0	49.0	22.8	12.1	-	-	-	-	-	-	-	-	-
QDTrack [66]	30.0	50.5	27.4	12.1	-	-	-	-	-	-	-	-	-
TETer [48]	33.3	51.6	35.0	13.2	-	-	-	-	-	-	-	-	-
OVTrack† [49]	34.7	49.3	36.7	18.1	-	-	-	-	-	-	-	-	-
STCN Tracker† [3]	-	-	-	-	5.5	0.9	17.5	0.7	2.5	0.6	-	-	-
Box Tracker† [3]	-	-	-	-	8.2	1.4	27.0	3.0	3.6	0.9	-	-	-
Detic [117]-SORT† [6]	-	-	-	-	-	-	-	-	-	-	12.8	21.1	6.6
Detic [117]-XMem †[16]	-	-	-	-	-	-	-	-	-	-	16.3	24.1	10.6
OV2Seg-R50† [83]	-	-	-	-	-	3.7	-	-	-	-	14.2	17.2	11.9
OV2Seg-B† [83]	-	-	-	-	-	4.9	-	-	-	-	21.1	27.5	16.3
UNINEXT (R50) [100]	31.9	43.3	35.5	17.1	-	-	-	-	-	-	-	-	-
GLEE-Lite†	40.1	56.3	39.9	24.1	22.6	12.6	36.4	18.9	19.1	11.0	19.6	22.1	17.7
GLEE-Plus†	41.5	52.9	40.9	**30.8**	26.9	17.2	38.8	23.7	23.9	15.5	**30.3**	**31.6**	**29.3**
GLEE-Pro†	**47.2**	**66.2**	**46.2**	29.1	**31.2**	**19.2**	**48.7**	**24.8**	**26.9**	**17.7**	23.9	24.6	23.3

Table 2. Comparison of GLEE to recent specialist and generalist models on object-level video tasks in a zero-shot manner. Evaluation metrics of BURST are reported separately for 'common', 'uncommon' and 'all' classes. The mAP computes mask IoU at the track level, HOTA is a balance of per-frame detection accuracy (DetA) and temporal association accuracy (AssA), and TETA that deconstructs detection into localization and classification components. The AP, AP_b, and AP_n in LV-VIS mean the average precision of overall categories, base categories, and novel categories. † does not use videos for training. The under-performance of Pro relative to Plus on LV-VIS is due to Pro employing larger training and inference resolutions, which prove to be sub-optimal for this specific dataset.

its annotations of over 1,200 object categories, showcasing a long-tail distribution. This distinction makes LVIS more representative of challenging real-world scenarios due to its broader category coverage. As indicated in Table 1, our model outperforms all generalist models on both COCO and LVIS benchmarks. Even when compared to SOTA specialist approaches, which are tailored with specific designs, our model remains highly competitive. This demonstrates that GLEE, while mastering universal and general object representations, concurrently maintains advanced performance. This characteristic is vitally important for adapting to a broad spectrum of downstream tasks requiring precise object localization. For the REC and RES tasks, we evaluated our model on RefCOCO [108], RefCOCO+ [108], and RefCOCOg [64], as show in Table 1, GLEE achieved comparable results with SOTA specialist methods PolyFormer [55], demonstrating strong capability to comprehend textual descriptions and showcasing potential to adapt to a broader range of multi-modal downstream tasks. In open-world instance segmentation tasks, GLEE outperforms previous arts ODISE [95] by 8.9 points, demonstrating the capability of identifying all plausible instance in an open-world scenario.

4.3. Zero-shot Evaluation Across Tasks

Zero-shot Transfer to Video Tasks. The proposed GLEE is capable of adapting to new data and even new tasks in a zero-shot manner, without the need for additional fine-tuning. We evaluate its zero-shot capability on three large-scale, large-vocabulary open-world video tracking datasets: TAO [19], BURST [3], and LV-VIS [83]. TAO comprises 2,907 high-resolution videos across 833 categories. BURST builds upon TAO, encompassing 425 base categories and 57 novel categories. LV-VIS offers 4,828 videos within 1,196 well-defined object categories. These three benchmarks require the model to detect, classify, and track all objects in videos, while BURST and LV-VIS additionally require segmentation results from the model. In Table 2, we compare the performance of our proposed model with existing specialist models. **Notably, the GLEE here is from the second training stage, which has not been exposed to images from these three datasets nor trained on video-level data.** Despite these constraints, GLEE achieves state-of-the-art performance that significantly exceeds existing methodologies. Specifically, GLEE surpasses the previous best method OVTrack by 36.0% in TAO, nearly triples the performance of the best baseline in BURST, and outperforms OV2Seg [83] by 43.6% in LV-VIS. This outstanding performance strongly validates the exceptional generalization and zero-shot capabilities of GLEE in handling object-level tasks across a range of benchmarks and tasks. It can be observed that GLEE yields more impressive results on video tasks, since the image tasks have progressed with plentiful data and models from lower costs, while video tasks have not due to higher costs. The models trained on extensive image data with strong general perception capabilities can effectively transfer to video tasks.

Method	Backbone	YTVIS 2019 val [102]			OVIS val [68]		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
SeqFormer [90]		47.4	69.8	51.8	15.1	31.9	13.8
IDOL [92]		49.5	74.0	52.9	30.2	51.3	30.0
VITA [33]		49.8	72.6	54.5	19.6	41.2	17.4
GenVIS [34]	ResNet-50	51.3	72.0	57.8	34.5	59.4	35.0
DVIS [112]		52.6	76.5	58.2	34.1	59.8	32.3
NOVIS [61]		52.8	75.7	56.9	32.7	56.2	32.6
UNINEXT [100]		53.0	75.2	59.1	34.0	55.5	35.6
GLEE-Lite		53.1	74.0	59.3	27.1/32.3	45.4/52.2	26.3/33.7
SeqFormer [90]		59.3	82.1	66.4	-	-	-
VITA [33]		63.0	86.9	67.9	27.7	51.9	24.9
IDOL [92]		64.3	87.5	71.0	42.6	65.7	45.2
GenVIS [34]	Swin-L	63.8	85.7	68.5	45.4	69.2	47.8
DVIS [112]		64.9	**88.0**	72.7	49.9	**75.9**	53.0
NOVIS [61]		65.7	87.8	72.2	43.5	68.3	43.8
GLEE-Plus		63.6	85.2	70.5	29.6/40.3	50.3/63.8	28.9/39.8
UNINEXT [100]	ConvNeXt-L	64.3	87.2	71.7	41.1	65.8	42.0
UNINEXT [100]	ViT-H	66.9	87.5	**75.1**	49.0	72.5	52.2
GLEE-Pro	EVA02-L	**67.4**	87.1	74.1	38.7/**50.4**	59.4/71.4	39.7/**55.5**

Table 3. Performance comparison on video instance segmentation tasks. (_/_) reports results from zero-shot and after fine-tuning.

Images	Method	AP	AP-categ	AP-descr	AP-descr-pos	AP-descr-S	AP-descr-M	AP-descr-L
All	RegionCLIP [114]	2.7	2.7	2.6	3.2	3.6	2.7	2.3
	Detic [117]	8.0	15.6	5.4	8.0	5.7	5.4	6.2
	MDETR [38]	-	-	4.7	9.1	6.4	4.6	4.0
	GLIP-T [47]	19.3	23.6	16.4	25.8	29.4	14.8	8.2
	GLIP-L [47]	25.8	32.9	21.2	33.2	37.7	18.9	10.8
	FIBER-B [23]	25.7	30.3	22.3	34.8	38.6	19.5	12.4
	GLEE-Lite	20.3	37.5	14.0	19.1	23.0	12.7	10.0
	GLEE-Lite-Scale	22.7	35.5	16.7	22.3	33.7	14.3	10.2
	GLEE-Plus	25.4	46.7	17.5	23.9	28.4	16.3	12.5
	GLEE-Plus-Scale	27.0	44.5	19.4	25.9	36.0	17.2	12.4

Table 4. Evaluation on the OmniLabel benchmark. The final AP value is the geometric mean of categories (AP-categ) and free-form descriptions (AP-descr).

We additionally provide performance comparison on classic video instance segmentation tasks, whose data is incorporated as image-level data during the second stage of training. As shown in Table 3, on the YTVIS2019 [102] benchmark, our model achieves SOTA results across various model sizes, surpassing all specialist models with complex designs to enhance temporal capabilities and the video unified model UNINEXT [100]. On the OVIS [68] benchmark, which features lengthy videos with extensive object occlusions where temporal capabilities of object features are particularly crucial, our model does not directly reach SOTA. However, after a few hours of simple fine-tuning, it still achieves SOTA performance. More details on VOS, RVOS and demonstrations of interactive segmentation and tracking can be found in supplementary materials.

Zero-shot Transfer to Real-world Downstream Tasks. To measure generalization on real-world tasks, we evaluate zero-shot performance on OmniLabel [76], which is a benchmark for evaluating language-based object detectors and encourages diverse free-form text descriptions of objects. As show in Table 4, compared to language-based detectors trained on large-scale caption data, GLEE significantly outperforms previous models in AP-categ. Due to the limited captions in our training dataset, it scores lower in AP-descr. By incorporating a more diverse set of box-caption data from the GRIT [67] to scale up our training set, the AP-descr can be elevated to a level comparable with existing models. A more comprehensive report on the zero-shot and few-shot performance on ODinW [44] and ablation studies are provided in the supplementary materials.

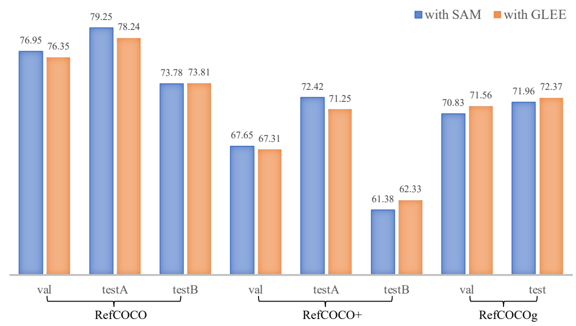

Figure 4. The performance comparison of replacing SAM with GLEE in LISA, GLEE achieves the same effectiveness as SAM in extracting objects.

4.4. Serve as Foundation Model

To explore whether GLEE can serve as a foundation model for other architectures, we selected LISA [43] for analysis, a mVLLM that combines LLAVA [54] with SAM [39] for reasoning segmentation. We substituted its vision backbone with a frozen, pretrained GLEE-Plus and fed the object queries from GLEE into LLAVA and remove decoder of LISA. We directly dot product the output SEG tokens with GLEE feature map to generate masks. As shown in Figure 4, after training for the same number of steps, our modified LISA-GLEE achieved comparable results to the original version, demonstrating the versatility of representations from GLEE and its effectiveness in serving other models.

5. Conclusion

We introduce GLEE, a cutting-edge object-level foundation model designed to be directly applicable to a wide range of object-level image and video tasks. Crafted with a unified learning paradigm, GLEE learns from diverse data sources with varying levels of supervisions. GLEE achieves state-of-the-art performance on numerous object-level tasks and excels in zero-shot transfer to new data and tasks, showing its exceptional versatility and generalization abilities. Additionally, GLEE provides general visual object-level information, which is currently missing in modern LLMs, establishing a robust foundation for object-centric mLLMs.

6. Acknowlegement

This research is supported by NSFC (No.62225603).

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. 2018. 2

[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. 2

[3] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1674–1683, 2023. 7

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 4

[5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 7

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 7

[7] 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. 1

[8] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 105–112. IEEE, 2001. 5

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[11] 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 IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[12] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5230–5238, 2017. 2, 5

[13] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[14] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 3, 6

[15] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 4, 6

[16] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2, 7

[17] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[19] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 7

[20] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 2, 4

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[22] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2302.01872*, 2023. 5

[23] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *NeurIPS*, 35:32942–32956, 2022. 8

[24] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 6

[25] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2

[26] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated

images. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 701–717, 2022. 3

[27] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 3

[28] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 3

[29] 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. 2, 4, 6

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[31] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2

[33] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *Advances in Neural Information Processing Systems*, 2022. 8

[34] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14623–14632, 2023. 2, 4, 8

[35] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 2

[36] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 5

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 3

[38] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 6, 8

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 2, 6, 8

[40] 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. 2, 6

[41] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[42] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2, 6

[43] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 8

[44] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 8

[45] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023. 3, 4, 6

[46] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, 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*, pages 2691–2700, 2023. 3, 6

[47] 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 *CVPR*, pages 10965–10975, 2022. 3, 8

[48] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*, 2022. 7

[49] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 7

[50] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 4, 6

[51] 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. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[52] 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. 2, 4, 6

[53] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 8

[55] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 2, 4, 6, 7

[56] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 5

[57] 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. *arXiv preprint arXiv:2103.14030*, 2021. 6

[58] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[59] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *arXiv preprint arXiv:2310.16667*, 2023. 3

[60] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2, 4

[61] Tim Meinhardt, Matt Feiszli, Yuchen Fan, Laura Leal-Taixe, and Rakesh Ranjan. Novis: A case for end-to-end near-online video instance segmentation. *arXiv preprint arXiv:2308.15266*, 2023. 8

[62] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 2016. 5

[63] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, page 728–755, 2022. 3

[64] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2, 6, 7

[65] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

[66] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021. 7

[67] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 6, 8

[68] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, pages 1–18, 2022. 4, 5, 6, 8

[69] 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. 1, 2, 3, 6

[70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 2, 3

[71] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[72] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[73] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[74] 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. 2

[75] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314, 2004. 5

[76] Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, and Dimitris Metaxas. Omnilabel: A challenging benchmark for language-based object detection. In *ICCV*, 2023. 8

[77] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 2, 4, 6

[78] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 3

[79] 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. 2, 4, 6

[80] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 5

[81] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 2

[82] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[83] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4057–4066, 2023. 7

[84] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3, 6

[85] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 6

[86] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[87] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[88] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 7

[89] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 256–263, 2014. 5

[90] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 2, 8

[91] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2, 4

[92] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, pages 588–605. Springer, 2022. 2, 4, 5, 8

[93] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2538–2550, 2023. 2

[94] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023. 6

[95] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 6, 7

[96] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S. Huang. Youtubevis dataset 2021 version. https://youtube-vos.org/dataset/vis/. 6

[97] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 5

[98] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2, 5

[99] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 2

[100] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3, 4, 6, 7, 8

[101] Haosen Yang, Chuofan Ma, Bin Wen, Yi Jiang, Zehuan Yuan, and Xiatian Zhu. Recognize any regions. 2023. 3

[102] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 4, 6, 8

[103] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 3, 6

[104] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang

Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*, 2022. 4

[105] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 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*, pages 23497–23506, 2023. 3

[106] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[107] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 3

[108] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 4, 6, 7

[109] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2, 3

[110] 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. 3

[111] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, pages 36067–36080, 2022. 6

[112] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1282–1291, 2023. 8

[113] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 2, 4

[114] 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. In *CVPR*, pages 16793–16803, 2022. 8

[115] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3

[116] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 2, 4

[117] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 7, 8

[118] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 2, 4, 6

[119] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 3

[120] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 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*, pages 16804–16815, 2022. 3

[121] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 3, 6

[122] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 2, 3