Scaling up Instance Segmentation using Approximately Localized Phrases

Karan Desai [1]
Ishan Misra [2]
Justin Johnson [1,2]
Laurens van der Maaten [2]

[1] University of Michigan
[2] Meta AI

Abstract

Training object detectors to segment large numbers of classes is challenging because they require training masks for each class. A potential solution is to partially supervise detectors using only bounding boxes for new object classes. While such boxes are easier to collect than masks, collecting them still requires cumbersome, exhaustive instance labeling from a pre-defined class ontology. We explore using natural language phrases for which a rough localization in the image is available; we refer to such weak supervision as *approximately localized phrases* (*ALP*s). We train detectors using masks from COCO dataset and learn to segment 300 Open Images classes, 240 of which do not have any labeled masks/boxes. Results show that ALP-supervised models outperform models that only train with masks for base classes. We also develop a simple one-stage detector to effectively learn from noisy localization of ALPs. Our model outperforms a comparable Mask R-CNN baseline when trained with ALPs. Taken together, our results suggest ALPs may be suitable for learning to segment a large number of object classes.

1 Introduction

Modern detectors [10, 15, 22, 31, 43, 56, 59] learn to segment objects using large image datasets with mask annotations [42, 46]. Training detectors to segment many more objects is difficult since collecting mask annotations is resource intensive [26]. Current scaling approaches for instance segmentation include *partially supervised* methods that use masks for few *base* classes and boxes for *novel* classes [35], and *weakly supervised* methods that replace masks *entirely* with boxes [34, 58]. Can we scale further using cheaper annotations?

While boxes are faster to annotate than masks [56], collecting detection labels is also cumbersome. Annotators need to exhaustively *spot* all instances belonging to a label ontology – either predefined [20, 42, 46] or gradually expanded [26]. Label ontologies vary across datasets and cannot handle linguistic variations like *synonyms* (*couch* $\Leftrightarrow$ *sofa*) or *hypernyms* (*rose* $\Leftrightarrow$ *flower*). Moreover, label collection requires strong inter-annotator agreement, so masks/boxes cannot be collected before finalizing labels to avoid duplication. For COCO, label collection accounted for $\approx$35% of the total cost of collecting mask annotations [18, 46].

To break away from label ontologies, there is a growing interest in using language supervision to pre-train vision models [8, 18, 19, 37, 60, 65]. However, the use of such supervision

*Correspondence to: kdexd@umich.edu. Work done by KD during an internship at Meta.

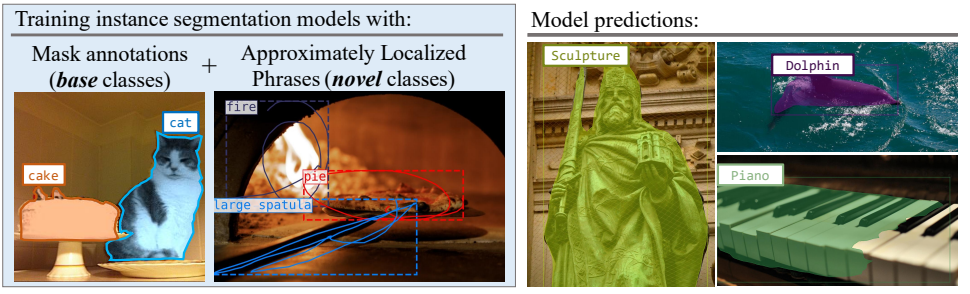

Figure 1: **Scaling up instance segmentation using ALPs. Left:** We train instance segmentation models using annotated masks and free-form phrases paired with boxes derived from mouse scribbles. **Right:** Our model can predict reasonably accurate masks for *novel* objects that do not have *any annotated masks or boxes in training data.*

for training detectors is hitherto limited. Natural language supervision has enabled *open-vocabulary detection* [75], which involves generalizing to *novel* classes without knowing them beforehand. One recent approach, ViLD [25], distills knowledge from CLIP/ALIGN models [37, 60]. However, this approach trains with labeled masks for *base* classes only.

In this paper, we explore using language supervision for learning to segment novel classes. Our task setup is straightforward, as shown in Fig. 1 (left) – we use training masks for some *base* classes and free-form phrases for *novel* classes. We also relax the need of *precise* bounding boxes by using only a rough localization through mouse scribbles/clicks. We call these **ALPs**, short for *Approximately Localized Phrases*. We curate ALPs from the recent Localized Narratives dataset (LocNar [59]). We use scribbles as an intermediary to derive boxes, so ALPs do not require special model components to process scribbles [45] or points [13, 16, 73]. Moreover, ALPs may be scaled up using image-text pairs are abundant on the web, along with boxes obtained via class-activation maps [55, 64, 78] or attention maps from large pre-trained vision models [37, 60, 74].

ALPs contain *imprecise* boxes that often do not completely enclose objects. This causes difficulty in training R-CNN detectors, as they draw supervision by matching anchors with ground-truth boxes as per area overlap. We think that boxes with imprecise edges should still have their centers close to true object centers. Based on this intuition, we extend FCOS [67] that uses center-based training supervision. We develop **FCOS-MO**: an **FCOS** with a **M**ask head and **O**pen-vocabulary classifier, to effectively handle noisy supervision of ALPs.

In summary, we study whether coarsely grounded language can aid in scaling instance-segmentation models to *novel* classes. In our experiments, we train FCOS-MO and Mask R-CNN baselines using COCO masks and LocNar ALPs and scale to 300 classes of Open Images [9, 42]. We observe that ALP-supervised FCOS-MO improves over an *open-vocabulary* baseline that only trains with *base* classes.

2 Related work

Partially box-supervised instance segmentation scales to many object classes by using training masks for a small subset and boxes for the rest. Hu *et al.* [35] formally define this task and introduce a model that generates *class-specific* mask head weights from box head via a hypernetwork [28]. Recent approaches simply use a *class-agnostic* mask head and learn shape priors [41], use foreground cues [5], or add auxiliary modules like mask

boundary prediction [21]. Birodkar *et al*. [6] simplify prior methods by showing that deep Hourglass [54] mask heads can seamlessly generalize to new object classes. Our task framing extends partially box-supervised instance segmentation – we aim to reduce its scaling cost by replacing precisely labeled boxes with ALPs.

Open-vocabulary detection generalizes to *novel* object classes without knowing them beforehand. Zareian *et al*. [75] propose this task as a simplification of zero-shot detection [3], and learn novel classes by pre-training the detector backbone on image-text pairs. ViLD [25] performs knowledge distillation [32] from CLIP/ALIGN [37, 60] to an R-CNN based detector. Recent works have also used pre-trained models to generate pseudo boxes for *novel* classes [23, 36]. These methods use training masks/boxes for *base* classes and rely on external models to achieve generalization. Detic [80] and MosaicOS [76] additionally use labeled images. Instead we use natural language annotations for novel classes.

Language supervised localization uses image-text data for learning representations that transfer to detection and segmentation tasks. VirTex [18] pre-trains the detector backbone on COCO Captions [14], LocTex [50] pre-trains on Localized Narratives [59]. MDETR [38] uses paired image-caption-boxes to train a text-modulated object detector. All these methods use training masks/boxes for object classes in downstream dataset. While pre-training may be beneficial, we focus on directly co-training with masks and natural language supervision.

Weakly supervised instance segmentation methods replace masks *entirely* with cheaper annotations, usually bounding boxes. SDI [39] uses GrabCut [63] to generate pseudo-masks and BBTP uses a prior that bounding boxes must tightly enclose masks. BoxInst [68] adds auxiliary losses based on mask shape and color similarity inside boxes. PointSup [46] proposes using a set of foreground-background annotated points together with boxes. While we also aim to use less masks, we emphasize on learning to segment *novel* classes.

Scribbles/clicks as localization cues are widely considered as an annotator-friendly alternative to masks/boxes. Early works used scribbles for interactive segmentation [7, 44] and learning semantic segmentation [45]. Scribbles and mouse clicks have aided scalable collection of mask annotations in datasets like COCO-stuff [9] and Open Images. Localized Narratives [42] used scribbles to ground speech and text in images. Recent works have used Localized Narratives to frame novel tasks using scribbles like scribble-guided image captioning [52] and retrieval [42]. In the similar spirit, we use scribbles to derive *approximate* boxes that can scale instance segmentation models to hundreds of *novel* classes.

3 Approach

We are interested in learning to segment a large set of object classes $\mathcal{C}$, without having access to training masks for all of them. We use training masks for some *base* classes $\mathcal{B}$, together with *approximately localized phrases* (ALPs) that cover the remaining *novel* classes $\mathcal{N}$. With this framing, we have $\mathcal{C} = \mathcal{B} \cup \mathcal{N}$. Current methods for scaling instance segmentation that use masks for *base* classes, and different types of supervision for *novel* classes are:

– **Open-vocabulary methods** use no explicit localized annotations for $\mathcal{N}$. They acquire *novel* classes by pre-training with image-caption data (OVR-CNN [75]) or by using external models like CLIP/ALIGN [37, 60] (ViLD [25]).
– **Label-supervised methods** use image-level labels for $\mathcal{N}$. Notable recent methods include Detic [80] and MosaicOS [76] that train detectors for ImageNet classes. These methods rely on image datasets that are collected using a pre-defined class ontology.

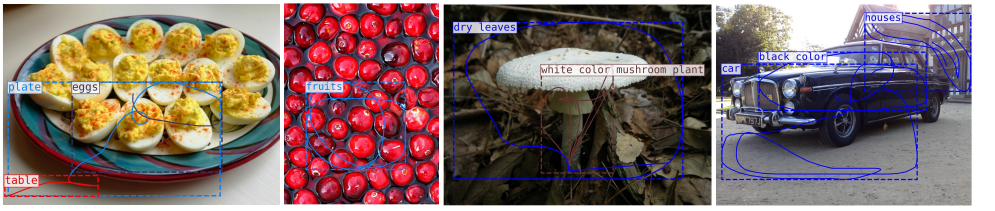

Figure 2: **ALPs from Localized Narratives:** LocNar annotators describe images verbally while moving the mouse over approximate image regions. We collect phrases paired with mouse scribbles; selected examples from LocNar-OID subset are shown above. ALP boxes are *imprecise* and often do not tightly enclose the underlying objects.

– **Box-supervised methods** [6, 35] use labeled boxes for $\mathcal{N}$. They are usually called *partially supervised* – we call them box-supervised to avoid ambiguity with ALP supervision.

Our task setup uses ALPs for $\mathcal{N}$, comprising free-form phrases and *approximate* boxes. In this section, we describe how we curate ALPs (Sec. 3.1), and introduce a simple FCOS-based detector to effectively learn from the noisy supervision of ALPs (Sec. 3.2).

3.1 Curating ALPs from captions and scribbles

We instantiate our task setup by using training masks from COCO [46], and curating ALPs from the Localized Narratives [59] dataset (LocNar). LocNar annotators verbally described images while moving their mouse over matching image regions. Image captions are speech transcriptions with each word having an utterance time-interval (t_1, t_2). They are paired with scribbles that are provided as a list of pixel coordinates and timesteps (x, y, t).

We perform parts-of-speech tagging on captions using RoBERTa-base [49] model from SpaCy [33] (`en-core-web-trf`). We extract *adjectives* and (consecutive) *nouns* as phrases. We discard phrases with *stuff* classes [9, 40] like *sky*, *water* as they are not suited for instance segmentation. For a particular phrase, let the union of utterance time-intervals of its words be (T_1, T_2). We extract the corresponding scribble segment as the set of points with a timestep in $(T_1 - t_p, T_2 + t_p)$, where t_p indicates a *temporal padding* hyperparameter [12].

We set $t_p = 0.5$ and collect ALPs from two subsets of LocNar: COCO and Open Images [42] (OID). The obtained phrase vocabulary is orders of magnitude larger than label ontologies of existing segmentation datasets [26, 42]. See Fig. 2 for few examples, and Supplementary for more details.

LocNar Subset ⇒	OID	COCO
Number of images	504K	118K
Number of ALPs	2.75M	825K
Unique phrases	51K	27.5K

Table 1: Counts of curated ALPs from the Localized Narratives dataset.

3.2 Model: FCOS-MO

Instance segmentation involves three subtasks – *object classification*, *box regression*, and *mask prediction*. All three are challenging when training with ALPs because:

– Phrases are highly varied, including synonyms (*couch* ⇔ *sofa*) and hypernyms (*sombrero* ⇔ *hat*). Some fine-grained classes may be entirely missing in training corpus.
– Boxes are *imprecise* and may give improper training supervision.
– Mask prediction must generalize to novel classes without training masks.

These challenges position our setup at the confluence of two tasks that have hitherto been studied in isolation: *viz.*, partially supervised instance segmentation [35] and open-vocabulary

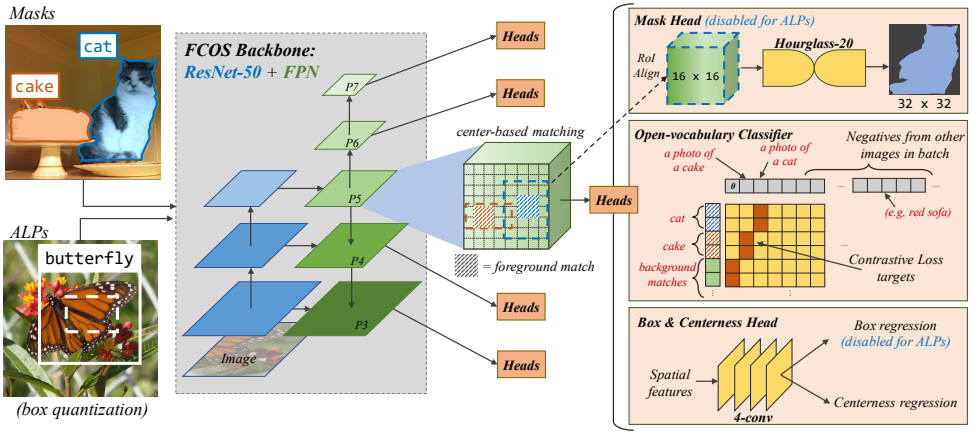

Figure 3: **FCOS-MO overview:** The backbone extracts *dense features* (**left**), performs *center-based* matches with GT boxes (**middle**) and passes them to heads (**right**). Mask head trains on RoI-aligned features from GT boxes, and is disabled for ALPs. Open-vocabulary classifier performs contrastive learning with *spatial features* from FPN and phrase features from CLIP. We perform box quantization augmentation with ALP boxes and disable box regression for them. Open-vocabulary classifier also has a *4-conv* stem, excluded for brevity.

detection [75]. State-of-the-art models for both these tasks [6, 25] build on R-CNN [31, 61], a two-stage anchor-based detector. Instead, we opt for one-stage anchor-free FCOS [67] that relies on box *centers* for training supervision rather than *edges* or *area*, which we think is better suited with imprecise boxes of ALPs. We call our model **FCOS-MO** – an **FCOS** model with a **M**ask head and **O**pen-vocabulary classifier to perform instance segmentation and handle free-form phrase supervision; see Fig. 3 for an overview.

Background: R-CNN and FCOS. In two-stage anchor-based R-CNN models, the first stage is a *region proposal network* (RPN) that uses fixed *anchor boxes* to predict some candidate proposal boxes that are likely to contain *any* object, and the second stage makes final predictions using these proposals. For training supervision, anchors are matched with GT boxes if they have high intersection-over-union (IoU). This matching gives correct supervision only when GT boxes have a proper area coverage over the underlying object. This is often not the case with boxes in ALPs, as seen in Fig. 2. Such noisy boxes may cause incorrect anchor matches during training, and hence may hurt model performance.

While box areas are heavily distorted in ALPs, we observe that their centers deviate less from true object centers. This makes FCOS favorable with ALPs as it relies on box centers for training supervision – FCOS matches *spatial features* with GT boxes based on proximity to box centers [67, 79], and includes centerness regression as an auxiliary training objective. This form of supervision can aid in learning to localize *novel* classes while using noisy boxes.

Mask head: Recent state-of-the-art approach in partially supervised instance segmentation, Deep-MARC [6], shows that a Mask R-CNN with a deep mask head (20+ layers) generalizes well to *novel* objects, without any bells and whistles. [6] also suggests training the mask head with *only* GT boxes, rather than with RPN proposals as is common in Mask R-CNN training. We adopt the same approach as it does not require any anchors or region proposals, which makes it easy to incorporate in FCOS.

Specifically, we use a 20-layer Hourglass [54] mask head with FCOS. This head uses

GT boxes to crop 16×16 FPN features via RoI-Align [31] and predict masks for them. Our mask head integration into FCOS is simpler than that in other recent models. For example, CondInst [56] predicts instance-aware mask head weights using box head. CenterMask [43] trains an attention-based mask head using predicted boxes as proposals. Recent works also use custom formats to represent masks, like polar coordinates [72] or compact vectors [77]. In contrast, our design allows using any Mask R-CNN based mask head with FCOS.

Open-vocabulary classifier: FCOS uses $|\mathcal{B}|$ independent sigmoid classifiers to detect $|\mathcal{B}|$ object classes. Since we train with free-form phrases, instead, we follow open-vocabulary and label-supervised methods [23, 25, 75, 80] and use embeddings from a pre-trained language model. This allows our model to flexibly handle the variations in natural language phrases (e.g. synonyms and hypernyms). We use the text encoder from CLIP `ViT-B/32` model to extract phrase embeddings and keep them fixed during training, similar to [25, 80]. We use the averaged embeddings of 63 text prompts per phrase, as used by ViLD.

We train this classifier using a CLIP-style contrastive loss [27, 29] – we first project the spatial features from the stem of classification head using 1×1 convolution such that their feature dimension matches that of phrase embeddings. Then we compute pairwise cosine similarities between each spatial feature and some candidate phrase embeddings, and apply softmax operation with temperature $\tau = 0.01$. The candidates comprise one GT target, and other phrases in the batch serve as *negatives* for the contrastive loss.

For dense features that were assigned to *background* (no GT box), we use a fixed *zero*-vector as the background embedding. During training, most of the spatial features are *background*. They may dominate in the loss computation and destabilize training, hence we downscale their log-probabilities by $\alpha = 0.1$, following prior works that use α-balanced cross-entropy to mitigate class imbalance in object detection [11, 48, 75].

3.3 Training and Inference with FCOS-MO

Training: We train FCOS-MO with mixed datasets comprising mask annotations and ALPs. For mask annotations, we use the name of object class as a phrase. When training with different sized datasets (e.g. COCO and LocNar-OID = 118K + 504K images), we sample alternating batches from each dataset. The training loss has *four* components: mask prediction (binary cross entropy), box regression (GIoU [62]), centerness regression (binary cross entropy), and region-phrase contrastive loss. We introduce a simple augmentation to improve training of FCOS-MO in presence of noisy boxes of ALPs, described below.

Box quantization: As shown in Fig. 2, boxes derived from scribbles often do not fully cover the underlying objects. This may lead to noisy supervision for both, FCOS and R-CNN based detectors. To mitigate this, we *quantize* these boxes to an imaginary grid over image. We stretch box edges outwards to increase the likelihood of enclosing the entire object. The grid we use is scale-invariant – it has square cells of $(Q \cdot L/256)$ pixels, where L is the shorter image edge (in pixels) and Q is a hyperparameter. We randomly sample $Q \in \{\phi, 8, 16, 32\}$ for every instance; ϕ means no quantization. We apply this augmentation only for ALPs, and disable box regression loss for those instances.

Inference: To detect classes in $\mathcal{C}$, we use their phrase embeddings along with the background embedding. We compute cosine similarity between spatial features and phrase embeddings, divide by temperature τ, and apply a $(|\mathcal{C}|+1)$-way softmax. We calibrate these scores by taking geometric mean with predicted centerness. After applying class-specific NMS, we input the RoI-aligned features of predicted boxes into mask head and obtain mask predictions.

3.4 Implementation details

We re-implement FCOS and Hourglass mask heads using the Detectron2 [71] framework in PyTorch [52]. As shown in Fig. 3, we use a ResNet-50 [30] backbone with FPN [47] and 4-*conv* prediction heads applied on spatial features from five FPN levels (named P3∼P7). We initialize the ResNet-50 backbone with ImageNet-pretrained weights and train the entire model using SGD with momentum = 0.9, and weight decay = 10^{-4}. We use a batch size of 32, distributed across 8 V100 GPUs and train for $270K$ iterations. We use a maximum learning rate = 0.05, apply a linear warmup for first $1K$ iterations, anneal it to zero using a cosine schedule [51]. We use SyncBN [58] in backbone, FPN, and heads. For heads, we keep separate SyncBN statistics per FPN level [70]. We use large-scale jittering (LSJ [24]) augmentation that has significantly boosted Mask R-CNN performance – random rescaling the image by $(0.5, 2.0)\times$ and using a padded 1024×1024 crop as input. We use automatic mixed precision [53] as implemented in PyTorch for faster training.

FCOS [67] trains for $90K$ iterations without mask head, SyncBN, or LSJ. We verify our re-implementation on fully supervised COCO; see `val2017` AP_{box}/AP_{mask} on the right. Our training recipe performs on par with original model (**39.5** vs **38.9**). In particular, our mask head enables segmentation, and adding SyncBN and LSJ improves performance. Hence, we are ready for a fair comparison between our model and strong Mask R-CNN baselines in our experiments.

Model	AP_{box}	AP_{mask}
FCOS *(original)*	38.9	–
FCOS *(reproduced)*	39.5	–
+ mask head	39.6	34.8
+ 3× training	42.0	37.0
+ SyncBN, LSJ	**43.7**	**38.7**

Table 2: Re-implementing FCOS (and improvements) in Detectron2.

4 Main Experiments

Our goal is to study whether ALPs can aid in scaling instance-segmentation to *novel* classes without any labeled masks/boxes. To this end, we train detectors to segment $|\mathcal{C}| = 300$ classes of Open Images [4] by using training masks from COCO dataset (118K images) and ALPs from LocNar-OID subset (Sec. 3.1). Upon manual inspection, we found that 60 OID classes are covered in COCO, hence we have $|\mathcal{B}| = 60$ *base* classes and $|\mathcal{N}| = 240$ *novel* classes.

Evaluation: For these experiments, we evaluate according to the Open Images 2019/20 challenge [1, 2] and report Mask AP@IoU=0.5 (AP_{50}) using maximum 100 predictions per image, for validation ($12K$ images, $23K$ masks) and test ($40K$ images, $74K$ masks) splits. We report performance for *base* and *novel* classes separately as AP_{50}-*base* and AP_{50}-*novel*. The latter is especially challenging because the training data has no labeled masks/boxes for *novel* classes that are semantically far from COCO (e.g, *Handgun*, *Dolphin*, *Sculpture*).

We compare ALP-supervised models with open-vocabulary and box-supervised models to study how different forms of supervision aid in learning to segment *novel* classes. State-of-the-art models for these tasks are based on the Mask R-CNN [31] architecture. Hence we include a Mask R-CNN baseline with similar components as FCOS-MO, described below.

Baseline: Mask R-CNN⋆. We start with a standard Mask R-CNN with ResNet-50 backbone [30] and FPN [47] as implemented in Detectron2, and make three modifications for fair comparison with our FCOS-MO model: (1) We use a *class-agnostic* box head in the second stage for generalizing to *novel* classes. (2) We use Hourglass-20 mask head (instead of 4-*conv*). (3) We replace the classification head in second stage with CLIP-based open-vocabulary classifier, similar to our FCOS-MO (Sec. 3.2). With these modifications, we

Model ⇓ AP_{50} ⇒	OID-v6 val *base*	OID-v6 val *novel*	OID-v6 test *base*	OID-v6 test *novel*
Open-vocabulary: COCO masks only				
Mask R-CNN⋆ *(ViLD-text style [25])*	64.6	25.8	62.8	21.4
ALP-supervised models: COCO masks + LocNar-OID ALPs				
Mask R-CNN⋆ *(ALPs without boxes)*	60.9	25.7	59.8	21.5
Mask R-CNN⋆	59.5	32.0 +7.8	56.7	27.8 +6.4
FCOS-MO	60.3	33.5 +9.3	60.0	30.5 +9.1
Box-supervised (oracles): COCO masks + OID boxes				
Mask R-CNN⋆ *(Deep-MARC style [6])*	61.9	54.2	61.8	50.5
FCOS-MO	60.6	51.3	60.2	47.5

Table 3: **Different task setups for scaling instance-segmentation:** We train FCOS-MO and Mask R-CNN⋆ on COCO→OID setup using different annotations for *novel* classes. We evaluate on OID val/test splits. ALP-supervised models outperform an open-vocabulary baseline on *novel* classes (*middle* vs *top*, green numbers show improvements). FCOS-MO outperforms Mask R-CNN⋆ when trained using ALP supervision.

refer the final model as **Mask R-CNN⋆** to avoid confusion. We train two variants of this baseline using COCO masks and different types of annotations for OID images:

– **Open-vocabulary:** We only train with COCO masks. This model and training setup is similar to ViLD [25], recent SOTA in open-vocabulary detection. ViLD also performs knowledge distillation from CLIP image encoder to RoI-aligned region features. We omit this due to computational constraints; it is orthogonal to our contributions.

– **Box-supervised:** We use box annotations of LocNar-OID images (1.9M boxes, 300 classes). These boxes are labeled and drawn by annotators with high precision, hence this variant serves as an *oracle* for ALP-supervised models. This training setup and model architecture follows [6], recent SOTA in partially box-supervised instance segmentation.

Mask R-CNN⋆ with ALPs: This model uses phrases with open-vocabulary classifier and uses ALP boxes to train RPN and box regression head in the second stage.

We also include an ablation that discards boxes of ALPs and only uses phrases as image-level labels. Without boxes, this model uses ALPs *only* to train the second-stage classifier. This way of training with image-level labels is similar to Detic [80]. We compute cosine similarity between RoI-aligned features from the largest-area RPN proposal and all phrase embeddings in the batch. For K ground-truth phrases, the target is a K-hot vector with values $1/K$ [1]. We scale this loss component by $\lambda = 0.1$, following Detic. With this ablation, we can assess the utility of *approximate localization* supervision available with ALPs.

Training details for all models are same as Sec. 3.4. We calibrate classifier scores at test-time by taking geometric mean with RPN *objectness* score, following ViLD and Detic.

Test-time hyperparameters: FCOS-based detectors have not been thoroughly benchmarked with OID in existing literature. It is unclear whether the test-time hyperparameters used with COCO are suitable for OID. So we search for the NMS threshold $\in \{0.5, 0.4, 0.3, 0.2, 0.1\}$ using the validation split. We use NMS threshold = 0.2 and score threshold = 0.05; these parameters work best for all trained models.

[1] Detic uses binary cross entropy loss defined over all image labels. We cannot use it due to free-form phrases.

Figure 4: **Qualitative examples. Top:** ALP-supervised FCOS-MO can predict reasonably accurate segmentation masks for a variety of *novel* object classes despite using no labeled masks/boxes for them. **Bottom:** per-category performance of ALP-supervised FCOS-MO.

Results: We show results in Table 3. Models that use ALPs for training outperform open-vocabulary Mask R-CNN⋆ on AP_{50}-*novel* (*middle vs top*). This suggests that ALPs can aid in learning to segment novel classes beyond solely relying on external pre-trained models. Among models using ALPs, label-supervised Mask R-CNN⋆ underperforms the other two, showing the utility of boxes derived from scribbles despite being very noisy. FCOS-MO performs the best, which shows its effectiveness in handling noisy supervision of ALPs.

Qualitative results: In Fig. 4 (top) we show some predicted masks by our ALP-supervised FCOS-MO model. To better visualize predictions, we *pooled* predictions of a single object class across OID val split and sort them by confidence score [17]. More random examples are included in Supplementary. We find reasonably accurate predictions for *novel* classes. Majority of errors are due to misclassification rather than inaccurate masks, this suggests that existing mask annotated datasets may be sufficient to scale up instance segmentation.

Per-category Mask AP distribution: Fig. 4 (bottom) shows per-category performance of ALP-supervised FCOS-MO. The model performs best on *base* classes (pizza, zebra, mug). At the middle of this distribution lie somewhat rare objects whose semantic neighbors may lie in COCO (*e.g.* limousine → car). We find a cluster of *novel* classes where the model performs the poorest – objects that are too small (*e.g.* chopsticks, adhesive tape), or *part* of another whole object, where the latter is described by annotators (e.g. human ear → person). Covering these may require specific instructions for annotators.

5 Additional Experiments and Ablations

The COCO→OID experiments in preceeding section show the utility of ALPs in scaling instance-segmentation models. In this section, we present experiments with COCO [46] to better contextualize our work with prior evaluation benchmarks. We experiment with two settings from prior works using *partial* mask supervision for *base* classes:

– **COCO `VOC-masks`** [6, 35]: This setup splits $|\mathcal{C}| = 80$ COCO classes into $|\mathcal{B}| = 20$ *base* VOC classes [20], and $|\mathcal{N}| = 60$ *base* non-VOC classes.

COCO `VOC-masks` [,]: *20 base, 60 novel classes*

Model	Setup ⇓ AP_{50} ⇒	*base*	*novel*
Mask R-CNN⋆	*open-vocabulary*	46.5	11.0
Mask R-CNN⋆	*ALP-supervised*	49.4	13.7
FCOS-MO	*ALP-supervised*	49.1	17.5
Mask R-CNN⋆	*box-sup. oracle*	49.8	29.0

COCO `ZSD` [,]: *48 base, 17 novel classes*

Model	Setup ⇓ AP_{50} ⇒	*base*	*novel*
Mask R-CNN⋆	*open-vocabulary*	39.8	13.4
Mask R-CNN⋆	*ALP-supervised*	42.5	18.1
FCOS-MO	*ALP-supervised*	42.7	25.3
Mask R-CNN⋆	*box-sup. oracle*	43.4	36.0

Table 4: **COCO experiments:** We train models on two different COCO setups and report metrics on COCO `val2017` split. ALP-supervised models improve open-vocabulary baselines. When trained with ALPs, FCOS-MO outperforms Mask R-CNN⋆.

AP_{50} ⇒	*base*	*novel*
FCOS-MO	49.1	**17.5**
HG-20 → *4-conv*	**49.2**	16.8

AP_{50} ⇒	*base*	*novel*
FCOS-MO	**49.1**	**17.5**
– *box quantization*	47.7	16.7

AP_{50} ⇒	*base*	*novel*
FCOS-MO	**49.1**	**17.5**
– *centerness head*	47.0	16.5

Table 5: **FCOS-MO Ablations:** All models are trained as per `COCO VOC-masks` setup. We conduct ablations with mask head architecture, box quantization, and centerness branch in FCOS-MO. Disabling these components degrades performance, especially on *novel* classes.

– **COCO `ZSD`** [,]: This setup uses $|\mathcal{B}| = 48$ *base* classes, and $|\mathcal{N}| = 17$ *novel* classes, remaining classes are ignored during training and evaluation.

Model Comparisons: We train four models like (Sec. 4): open-vocabulary Mask R-CNN⋆ baseline, ALP-supervised Mask R-CNN⋆ and FCOS-MO (using LocNar-COCO ALPs; see Sec. 3.1), and box-supervised Mask R-CNN⋆ oracle. Note that Detic-style training is not compatible with this setup – in COCO, images with masks and ALPs overlap with images having ALPs. We keep all implementation details same, except training only for 90K iterations due to small dataset size (≈24 COCO epochs).

Results: Tab. 4 shows AP_{50}-*base* and AP_{50}-*novel* for both setups. We observe exactly same trends as Tab. 3. Notably the gap between ALP-supervised models and box-supervised oracle is larger – segmentation is challenging in these setups because of less training masks.

Ablations: We conduct basic ablations to study the effect of different modeling components in FCOS-MO. We select the ALP-supervised FCOS-MO trained on COCO `VOC-masks` setup (Tab. 4 left, third row) as a base model and train three separate ablations – (1) Replacing Hourglass-20 mask head with the standard 4-*conv* head of Mask R-CNN, (2) Training without box quantization for ALPs, to show how this augmentation makes our model less sensitive to noisy boxes, (3) Disabling the centerness branch to observe the effect of *only* center-based matching with ALPs. All results are shown in Tab. 5. Each of these ablations underperforms our full FCOS-MO model, especially for *novel* classes.

6 Conclusion

In summary, we study scaling instance-segmentation models using natural language that is approximately localized in images, and show a concrete instantiation of this task with existing datasets. We introduced ALPs as a scalable alternative to using *precisely* labeled masks and bounding boxes for scaling instance segmentation models. Our experiments show that ALPs can indeed be a possible alternative, surpassing open-vocabulary baselines and reaching towards the performance of oracle models that train on labeled bounding boxes. Future work could explore scaling up further by curating weakly aligned image-text pairs, with the help of large pre-trained vision models.

Acknowledgments

We thank Mohamed El Banani, Julius Berner, Bowen Cheng, Richard Higgins, Gaurav Kaul, Nilesh Kulkarni, Chris Rockwell, Steffen Schneider, Dandan Shan, Ramakrishna Vedantam, and Xingyi Zhou for helpful discussions. We thank all the anonymous reviewers for constructive feedback during the review phase.

References

[1] Open Images Challenge instance segmentation evaluation. https://storage.googleapis.com/openimages/web/evaluation.html. Accessed on Mar 1 2022. 7

[2] Open Images 2020 Instance Segmentation Track, RVC Challenge. https://kaggle.com/c/open-images-instance-segmentation-rvc-2020, 2020. 7

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-Shot Object Detection. In *ECCV*, 2018. 3, 10

[4] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 2, 7

[5] David Biertimpel, Sindi Shkodrani, Anil S. Baslamisli, and Nóra Baka. Prior to Segment: Foreground Cues for Weakly Annotated Classes in Partially Supervised Instance Segmentation. In *ICCV*, 2021. 2

[6] Vighnesh Birodkar, Zhichao Lu, Siyang Li, Vivek Rathod, and Jonathan Huang. The surprising impact of mask-head architecture on novel class segmentation. *arXiv preprint arXiv:2104.00613*, 2021. 3, 4, 5, 8, 9, 10

[7] Yuri Boykov and Marie-Pierre Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In *ICCV*, 2001. 3

[8] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. 1

[9] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*, 2018. 3, 4

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 6

[12] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *ICCV*, 2021. 3, 4

[13] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as Queries: Weakly Semi-supervised Object Detection by Points. In *CVPR*, 2021. 2

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[15] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. *arXiv*, 2021. 1

[16] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *CVPR*, 2022. 2, 3

[17] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021. 9

[18] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021. 1, 3

[19] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 1

[20] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2009. 1, 9

[21] Qi Fan, Lei Ke, Wenjie Pei, Chi-Keung Tang, and Yu-Wing Tai. Commonality-parsing network across shape and appearance for partially supervised instance segmentation. In *ECCV*, 2020. 3

[22] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 1

[23] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards Open Vocabulary Object Detection without Human-provided Bounding Boxes. *arXiv preprint arXiv:2111.09452*, 2021. 3, 6

[24] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 7

[25] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 2, 3, 5, 6, 8

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 4

[27] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 6

[28] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 2

[29] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 6

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 5, 6, 7

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531. 3

[33] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303. 4

[34] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In *NeurIPS*, 2019. 1

[35] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018. 1, 2, 4, 9, 10

[36] Dat T. Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling. *arXiv preprint arXiv:2111.12698*, 2021. 3

[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 *ICML*, 2021. 1, 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*, 2021. 3

[39] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 3

[40] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *CVPR*, 2019. 4

[41] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. ShapeMask: Learning to segment novel objects by refining shape priors. In *ICCV*, 2019. 2

[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. 1, 2, 4

[43] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020. 1, 6

[44] Anat Levin, Dani Lischinski, and Yair Weiss. A Closed-Form Solution to Natural Image Matting. *TPAMI*, 2008. 3

[45] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-supervised Convolutional Networks for Semantic Segmentation. In *CVPR*, 2016. 2, 3

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 4, 9

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7

[48] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[50] Zhijian Liu, Simon Stent, Jie Li, John Gideon, and Song Han. LocTex: Learning Data-Efficient Visual Representations from Localized Textual Supervision. In *ICCV*, 2021. 3

[51] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 7

[52] Zihang Meng, Licheng Yu, Ning Zhang, Tamara L Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In *CVPR*, 2021. 3

[53] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018. 7

[54] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 5

[55] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2

[56] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. 1

[57] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 7

[58] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. MegDet: A large mini-batch object detector. In *CVPR*, 2018. 7

[59] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 2, 3, 4

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5

[62] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection over Union. In *CVPR*, 2019. 6

[63] Carsten Rother, Vladimir Kolmogorov, , and Andrew Blake. GrabCut: interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, 2004. 3

[64] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *ICCV*, 2017. 2

[65] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model. In *CVPR*, 2022. 1

[66] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional Convolutions for Instance Segmentation. In *ECCV*, 2020. 1, 6

[67] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A Simple and Strong Anchor-Free Object Detector. *TPAMI*, 2020. 2, 5, 7

[68] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *CVPR*, 2021. 1, 3

[69] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 1

[70] Yuxin Wu and Justin Johnson. Rethinking" batch" in batchnorm. *arXiv preprint arXiv:2105.07576*, 2021. 7

[71] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[72] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020. 6

[73] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. RepPoints: Point Set Representation for Object Detection. In *ICCV*, 2019. 2

[74] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[75] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2, 3, 5, 6, 10

[76] Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. Mosaicos: a simple and effective use of object-centric images for long-tailed object detection. In *ICCV*, 2021. 3

[77] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *CVPR*, 2020. 6

[78] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2

[79] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5

[80] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. 2022. 3, 6, 8