Mask-free OVIS: Open-Vocabulary Instance Segmentation without Manual Mask Annotations

Vibashan VS[⋆,∗], Ning Yu[†], Chen Xing[†], Can Qin[‡], Mingfei Gao[†],
Juan Carlos Niebles[†], Vishal M. Patel[⋆], Ran Xu[†]

[⋆]Johns Hopkins University, [‡]Northeastern University, [†]Salesforce Research

{vvishnu2, vpatel36}@jhu.edu, qin.ca@northeastern.edu,

{ning.yu, cxing, jniebles, ran.xu}@salesforce.com

Abstract

Existing instance segmentation models learn task-specific information using manual mask annotations from base (training) categories. These mask annotations require tremendous human effort, limiting the scalability to annotate novel (new) categories. To alleviate this problem, Open-Vocabulary (OV) methods leverage large-scale image-caption pairs and vision-language models to learn novel categories. In summary, an OV method learns task-specific information using strong supervision from base annotations and novel category information using weak supervision from image-captions pairs. This difference between strong and weak supervision leads to overfitting on base categories, resulting in poor generalization towards novel categories. In this work, we overcome this issue by learning both base and novel categories from pseudo-mask annotations generated by the vision-language model in a weakly supervised manner using our proposed Mask-free OVIS pipeline. Our method automatically generates pseudo-mask annotations by leveraging the localization ability of a pre-trained vision-language model for objects present in image-caption pairs. The generated pseudo-mask annotations are then used to supervise an instance segmentation model, freeing the entire pipeline from any labour-expensive instance-level annotations and overfitting. Our extensive experiments show that our method trained with just pseudo-masks significantly improves the mAP scores on the MS-COCO dataset and OpenImages dataset compared to the recent state-of-the-art methods trained with manual masks. Codes and models are provided in https://vibashan.github.io/ovis-web/.

1. Introduction

Instance segmentation is a challenging task as it requires models to detect objects in an image while also precisely
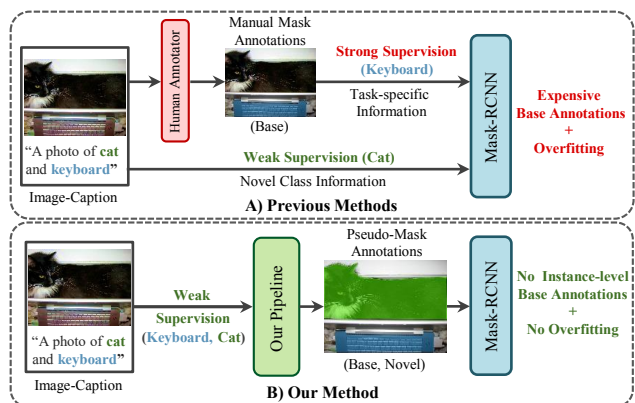

Figure 1. **A) Previous Methods:** Learn task-specific information (detection/segmentation) in a fully-supervised manner and novel category information with weak supervision. During training, this difference in strong and weak supervision signals leads to overfitting and requires expensive base annotations. **B) Our method:** Given image-caption pairs, we generate pseudo-annotations for both base and novel categories under weak supervision, solving the problems of labour-expensive annotation and overfitting.

segment each object at the pixel-level. Although the rise of deep neural networks has significantly boosted the state-of-the-art instance segmentation performance [7, 15, 42], these methods are still trained for a pre-defined set of object categories and are data-hungry [33]. Particularly, one needs to manually annotate thousands of instance-level masks for each object category, which takes around 78 seconds per instance mask [3]. If we look at this quantitatively on Open Images [21], a large-scale dataset with 2.1M instance-level mask annotations requires around 5 years of human labour. Even after extensive annotation, these training datasets are still limited to a small number of categories and segmenting objects from a novel category requires further annotation. Therefore, it is difficult to scale up existing methods to segment a large number of categories due to intensive labour.

Recently, Open-Vocabulary (OV) methods have gained much attention due to their success in detecting [11, 14,
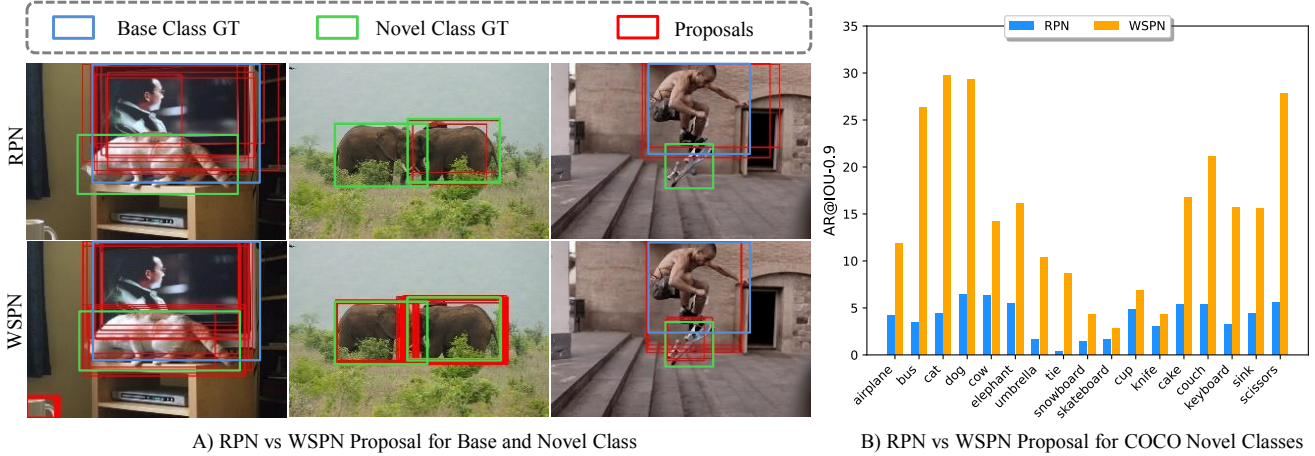
A) RPN vs WSPN Proposal for Base and Novel Class

B) RPN vs WSPN Proposal for COCO Novel Classes

Figure 2. RPN is supervised using bounding box annotations from COCO base and WSPN is supervised using image-labels from COCO base. **A)** WSPN produces better quality proposals for novel object categories compared to fully-supervised RPN. **B)** WSPN consistently produces better recall for all COCO novel categories than RPN.

46, 52] and segmenting [17] *novel* categories beyond *base* (training) categories. An OV method learns task-specific information (detection/segmentation) from base categories that have manual instance-level bounding boxes or masks and learns novel category information from the pre-trained Vision-Language Model (VLM) [19, 34] (see Fig. 1). All these methods produce promising results on novel categories by leveraging different forms of weak supervision such as caption pretaining [46, 50], knowledge distillation [14, 52] and pseudo-labelling [17, 28]. However, all above-mentioned OV methods still rely on the manually-annotated base categories to improve their performances on novel categories. Without fine-tuning on base categories, existing OV methods lack task/domain specific knowledge and the performances on novel categories will be affected [14, 46].

Although manual instance-level annotations of base categories are critical to open-vocabulary segmentation methods, we find empirically that such fully-supervised information causes OV methods to overfit to base categories, leading to a higher failure rate when evaluated on novel categories. Specifically, OV methods utilize a region proposal network (RPN) [37] supervised with bounding box annotations obtained from the base categories to generate a set of bounding box proposals for all objects categories in a given image [46]. The feature representation from these object proposals is later matched with text embedding to learn the visual-semantic space for base and novel categories [46]. Therefore, the quality of proposals generated for novel object categories plays a key role in determining the performance in later stages. However, from our experiments, we find that many objects of novel categories wouldn't be included in such proposals due to the RPN's overfitting to base categories. Fig. 2 (A) - Top gives some examples where the RPN trained with COCO base categories fails to generate high-quality region proposals for novel categories such as elephant, cat and skateboard. Therefore, a fully-

supervised proposal network is a bottleneck in OV pipeline due to its poor generalization towards novel categories.

Given the aforementioned observations of poor generalization, we raise the question of whether we can improve the generalization by using weak supervision instead of relying on strong supervision from manual base annotations. If so, we can reduce overfitting to the base categories and the requirement for costly instance-level human annotations can be entirely removed from the pipeline. Our preliminary experiments give us some hope. Our experiments show that if we train a weakly-supervised proposal network (WSPN) with image-level annotations instead of box-level annotations, the region proposals it generates can better generalize to novel objects. As shown in Fig. 2 A), the novel objects that the RPN proposals miss are covered by WSPN. Fig. 2 B) shows WSPN proposals have consistently better average recall than RPN for all COCO novel categories, indicating that WSPN proposals are more likely to cover the ground-truth bounding boxes of novel objects.

Inspired by these observations, we propose open-vocabulary segmentation without manual mask annotations. We do not use any human-provided box-level or pixel-level annotations during the training of our method. We first train a simple WSPN model with image-level annotations on base categories as a proposal generator to generate proposals for all objects given an image. Then, we adopt pre-trained vision-language models to select proposals as pseudo bounding boxes for novel objects. Given a novel object's text name, we can utilize the name as a text prompt to localize this object in an image with a pre-trained vision-language model. To obtain a more accurate pseudo-mask that covers the entire object, we conduct iterative masking with GradCAM [39] given the vision-language model. Finally, we train a weakly-supervised segmentation (WSS) [20] network with previously generated bounding box and GradCAM activation map to obtain pixel-level annotation.

Our contributions are summarized as follows: **(1)** We propose a Mask-free OVIS pipleine where we produce manual-effort-free pseudo-mask annotations for base and novel instance segmentation using open-vocabulary and weakly supervised techniques. **(2)** We propose a novel pseudo-mask generation pipeline leveraging a pre-trained vision-language model to generate instance-level annotations. **(3)** Benefiting from pseudo-labels, our method sets up SOTA's for both detection and instance segmentation tasks compared to recent methods trained with manual masks on MS-COCO and OpenImages datasets.

2. Related Work

Weakly-supervised learning (WSL) methods require less manual annotation effort and only rely on partial information about the task compared to fully-supervised methods. WSL applications include object detection [1, 16, 44, 47], segmentation [22,23,40], action recognition, etc [26,41,48]. The class activation map (CAM) [53] and grad-CAM [39] are popular approaches which exploits pre-trained network to generate saliency or activation map for target classes. Li *et al*. [26] enables weakly-supervised semantic segmentation via CAM-guided attention.

Vision-Language Models. With the rise of large-scale text-pretraining with attention-based models [9,49], vision-language models (VLM) have caught increasing attention due to their strong performance in downstream visual understanding tasks [27,31,34]. Early VLM models [27,31] applied the pre-trained models with constraints of the class set. With the help of contrastive learning, CLIP [34] has enabled VLM in the wild with large-scale multimodal learning upon the 400M noisy data crawled from the web. Several methods have extended CLIP for high efficiency model training and cycle consistency [13,24,25]. ALBEF [25] utilizes cross-modal attention between image patches and text, leading to more grounded vision and language representation learning. Compared to CLIP, ALBEF excels in object localization, while CLIP performs better at zero-shot classification due to its larger training dataset.

Open-Vocabulary. Open-Vocabulary methods [6, 10, 32, 36, 46, 50] scale up their vocabulary size for object detection or instance segmentation tasks by transferring knowledge from pre-trained vision-language models. Zareian *et al*. [46] proposed an open-vocabulary object detection method where a visual encoder is trained on image-caption pair to model object semantics and then transfer to zero-shot object detection. ViLD [14] and RegionCLIP [52] use pre-trained CLIP [34] to distill knowledge and enhance the vision-text embedding space effectively. Similarly, Huynh *et al*. [17] proposes a robust pseudo-labelling, for instance segmentation where a teacher model generates pseudo-label for novel categories and is later used to supervise a student model. Therefore, all these methods require manually-

annotated base categories to improve their performance on novel categories. In contrast, our method generates pseudo-labels for both base and novel categories by leveraging a pre-trained VLM, freeing the entire pipeline from human-provided instance-level annotations. Similar to our method, Gao *et al*. [11] generates pseudo-labels using pre-trained VLM, but their approach still uses a proposal generator that relies on human-provided bounding boxes for the base categories. Our method, on the other hand, requires no such annotations, making it completely free from any human-provided box-level or pixel-level annotations.

3. Mask-free OVIS

Our pipeline consists of two stages: (I) Pseudo-mask generation, and (II) Open-vocabulary instance segmentation. The goal of (I) is to generate box-level and pixel-level annotations by leveraging the region-text alignment property of a pre-trained vision-language model as illustarted in Fig. 3. In (II), we train a Mask-RCNN using the generated pseudo annotations.

3.1. Pseudo-Mask Generation

Region-Text Attention Scores. The inputs to the vision-language model are image $\mathbf{I}$ and caption $\mathbf{C} = \{c_1, c_2, ..., c_{N_c}\}$ pair, where N_c is the number of words in the caption (including [CLS] and [SEP]) [9, 25]. In a VLM, a text encoder is utilized to get text representations $\mathbf{T} \in \mathbb{R}^{N_C \times d}$ and an image encoder is utilized to extract region representation $\mathbf{R} \in \mathbb{R}^{N_R \times d}$, where N_R is the number of regions in the image. To fuse the information from both image and text encoders, a multi-modal encoder with M consecutive cross-attention layers is utilized [19, 25]. These cross-attention layers learn a good region-text alignment [25, 34] and are utilized to obtain regions corresponding to the object of interest c_t from the caption. In particular, for the m-th cross-attention layer, the visual region attention scores $\mathbf{X}_t^m$ for the object of interest c_t is calculated as:

$$\mathbf{X}_t^m = \text{Softmax}\left(\frac{\mathbf{h}_t^{m-1}\mathbf{R}^T}{\sqrt{d}}\right), \quad (1)$$

$$\mathbf{h}_t^n = \mathbf{X}_t^m \cdot \mathbf{R}. \quad (2)$$

where d is a scalar and $\mathbf{h}_t^{m-1}$ is the hidden representation obtained from the previous $(m-1)$-th cross-attention layer. **GradCAM Activation Map.** After obtaining attention scores $\mathbf{X}_t^m$, we employ Grad-CAM [39] to visualize the activated regions. As in [39], we take the image-caption similarity (S) output from the multi-modal encoder's final layer and calculate the gradient with respect to the attention scores. The activation map ϕ_t for object c_t is:

$$\phi_t = \mathbf{X}_t^m \cdot \max\left(\frac{\partial S}{\partial \mathbf{X}_t^m}, 0\right). \quad (3)$$

Iterative Masking. During VLM training, an object's most discriminative regions easily get aligned towards object text representation [25, 34]. As a result, ϕ_t is localized towards
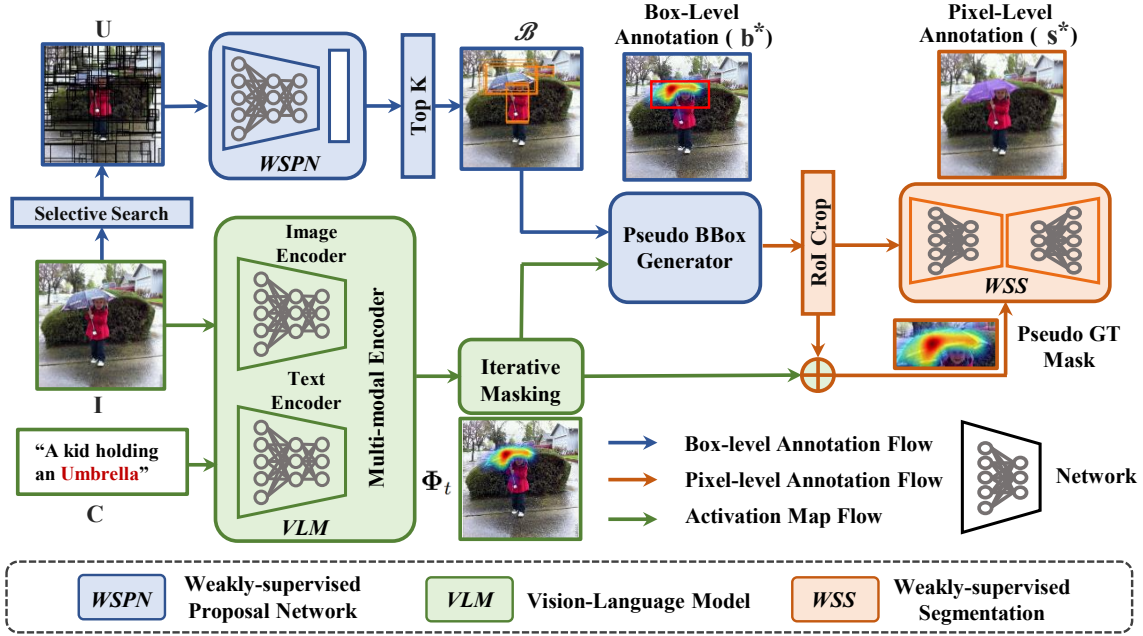
Figure 3. Illustrative overview of our pseudo-mask generation pipeline. Given an image-caption pair and pre-trained VLM, we generate an activation map for the object of interest ("umbrella") and enhance it using iterative masking strategy. We generate box-level annotations using an activation map as a guidance function to select the best WSPN proposals covering the object. We crop the image corresponding to the generated pseduo bounding box and perform weakly-supervised segmentation to obtain pixel-level annotations.

the most discriminative region and fails to cover the object completely [39]. However, when we mask out the most discriminative regions, GradCAM activations are shifted towards other discriminative regions (see Fig. 4). We propose a simple iterative masking strategy to obtain better activation where the most activated part of the object is replaced with image mean and the new activation map is computed following Eq. 1 and 3. The final activation map is:

$$\mathbf{\Phi}_t = \bigcup_{i=1}^{G} \mathcal{IM}(\phi_t^i) \qquad (4)$$

where G is a hyper-parameter indicating the number of masking iterations and $\mathcal{IM}(\cdot)$ normalize and threshold ϕ_t by 0.5. We utilize the activation map $\mathbf{\Phi}_t$ as a guidance function to generate box-level and pixel-level annotations.

Weakly-Supervised Proposal Network. To generate box-level annotations, we require bounding box proposals covering the activated region $\mathbf{\Phi}_t$. As explained in the introduction, the fully-supervised RPN network produces poor proposals for novel categories, making it less preferable. To overcome this, we propose a weakly-supervised proposal network (WSPN), which generates object proposals trained in a weakly-supervised manner. Given an Image $\mathbf{I}$, the WSPN is supervised with image-labels $\mathbf{Y} = \{y_1, y_2, ..., y_C\}$, where $y_c = 0$ or 1 indicates the absence or presence of class c in $\mathbf{I}$ and C denotes total number of classes [5]. First, we utilize Selective Search to generate a set of unsupervised proposal $\mathbf{U} = \{u_1, u_2, ..., u_N\}$ where N is the total number of bounding box proposals. Then, the Image $\mathbf{I}$ and proposal $\mathbf{U}$ are fed into a CNN backbone
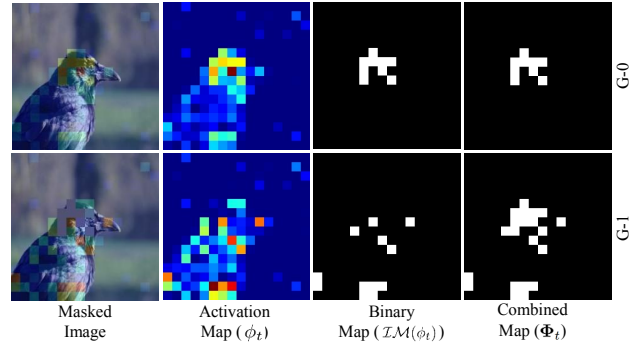

Figure 4. Comparison between activation map $\mathbf{\Phi}_t$ generated at G 0 and 1. For G=0, the most discriminative parts of an object (bird's head) gets activated (bird's head). After masking, for G=1 we can observe that the activation map has shifted to less discriminative part (bird's body). Thus, by combining activation from both steps, we obtain a better activation map trying to cover entire object $\mathbf{\Phi}_t$.

to extract features and RoI pooling layer [37] to obtain RoI pooled feature vectors. Following [5], the pooled feature vectors are passed to a classification and detection branch to generate two matrices $\mathbf{W}^{cls}, \mathbf{W}^{det} \in \mathbb{R}^{C \times N}$. Then, $\mathbf{W}^{cls}$ and $\mathbf{W}^{det}$ matrices are normalized along the category direction (column-wise) and proposal direction (row-wise) by the softmax layers $\sigma(\cdot)$ respectively. From $\mathbf{W}^{cls}$ and $\mathbf{W}^{det}$, the instance-level classification scores for object proposals are computed by the element-wise product $\mathbf{W_C} = \sigma(\mathbf{W}^{cls}) \odot \sigma(\mathbf{W}^{det})$. and image-level classification score for the c_{th} class is computed as $p_c = \sum_{i=1}^{N} w_{i,c}$. Following [38], we select the high-scoring proposals are used as

pseudo-regression targets ($\hat{\mathbf{T}} = \{\hat{t}(u_1), \hat{t}(u_2), ..., \hat{t}(u_N)\}$) for low-scoring proposals to make sure objects can be more tightly captured. To this end, the classification loss and regression loss for WSPN are calculated as:

$$\mathcal{L}_{wspn} = -\sum_{c=1}^{C} y_c \log p_c + (1 - y_c) \log(1 - p_c) +$$
$$\frac{1}{N} \sum_{u=1}^{N} \mathcal{L}_{smoothL1}(\hat{t}(u_i), u_i). \tag{5}$$

WSPN is trained to localize and classify objects by minimizing Eq. 5. The trained WSPN model is used to generate object proposals and top K proposal candidates over all the classes $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_K\}$ are selected by sorting proposal confidence scores obtained from $\mathbf{W}^{det}$. Finally, from the top proposal candidates $\mathcal{B}$, we select the proposal which overlaps the most with $\mathbf{\Phi}_t$ as pseudo box-bounding:

$$\mathbf{b}^* = \arg\max_{\mathbf{b} \in \mathcal{B}} \frac{\sum_{\mathbf{b}} \mathbf{\Phi}_t}{\sqrt{|\mathbf{b}|}}, \tag{6}$$

where $\mathbf{b}^*$ is the pseudo box-level annotation and $\sum_{\mathbf{b}} \mathbf{\Phi}_t$ indicates the summation of the activation map values within a box proposal $\mathbf{b}$, and $|\mathbf{b}|$ indicates the proposal area.

Weakly-Supervised Segmentation. Once we obtain the pseudo bounding box $\mathbf{b}^*$, we crop image $\mathbf{I}$ to obtain the corresponding image patch. The cropped patch is then fed into a simple three-layer CNN network to perform pixel-level segmentation. To supervise the CNN network, we generate pseudo ground-truth $\mathbf{\Theta}$ using $\mathbf{\Phi}_t$ and $\mathbf{b}^*$, where we sample Z points as foreground $F_z = \{f_i\}_{i=1,..,Z}$ and background $B_z = \{b_i\}_{i=1,..,Z}$ and each point is set to 1 or 0, respectively. Specifically, the foreground and background points are sampled from the most and least activated part of $\mathbf{\Phi}_t$ inside $\mathbf{b}^*$. The pseudo ground-truth $\mathbf{\Theta}$ is of size $\mathbf{b}^*$ and we supervise the network predictions only at sampled points. Thus, the segmentation loss obtained from these weak points is computed as follows:

$$\mathcal{L}_{wss} = \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(f_i), \mathbf{\Theta}(f_i)) + \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(b_i), \mathbf{\Theta}(b_i)), \tag{7}$$

where $\mathbf{s}^*$ is the pseudo pixel-level annotation of size $\mathbf{P}$ and $\mathcal{L}_{ce}$ indicates cross-entropy loss.

To this end, given an image $\mathbf{I}$ and caption $\mathbf{C}$ pair, we generate pseudo box-level $\mathbf{b}^*$ and pixel-level $\mathbf{s}^*$ annotation for the object of interest c_t. In practice, the pseudo-mask annotations are generated for a pre-defined set of object categories obtained from training vocabulary. Fig. 5 visualizes the generated pseudo-mask annotations for various categories. Activated regions correspond well with objects of interest and the generated pseudo-mask are of good quality.

3.2. Open-Vocabulary Instance Segmentation

After generating pseudo-mask annotations, we train an open-vocabulary instance segmentation model. Following

[46], we employ a `Mask-RCNN` as the instance segmentation model, where a class-agnostic mask head is utilized to segment objects and the classification head is replaced with embedding head h_{emb}. Given a Image $\mathbf{I}$, an encoder network extracts image features and region embeddings, $\mathbf{R} = \{\mathbf{r}_i\}_{i=1,...,N_r}$, are obtained by RoI align [15] followed by a fully connected layer, where N_r denotes the number of regions. The similarity between the region and text embedding pair is calculated as follows:

$$p(\mathbf{r}_i, \mathbf{c}_j) = \frac{\exp\left(h_{emb}(\mathbf{r}_i) \cdot \mathbf{c}_j\right)}{\exp(h_{emb}(\mathbf{r}_i) \cdot \mathbf{bg}) + \sum_k \exp\left(h_{emb}(\mathbf{r}_i) \cdot \mathbf{c}_k\right)}, \tag{8}$$

where, $\mathbf{C} = \{\mathbf{bg}, \mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{N_C}\}$, are object vocabulary text representation obtained from pre-trained text encoder, where N_C is the training object vocabulary size. To learn the semantic space, negative pairs are pushed away and positive pairs are pulled together using cross entropy loss obtained from $\mathbf{b}^*$. The class-agnostic mask head is supervised by minimizing standard segmentation loss [15] obtained from $\mathbf{s}^*$. During inference, the similarity between the region proposals embedding and text embedding from a group of object classes of interest is calculated. The region is then assigned to a class with the highest similarity.

4. Experiments and Results

Datasets. Following [46], we conduct experiments on MS-COCO [29] with data split of 48 base categories and 17 novel categories. The processed COCO dataset contains 107,761 training images and 4,836 test images. Following [17], we conduct experiments on Open Images [21] to verify the effectiveness of our method on the large-scale dataset. The Open Images dataset consists of 300 categories with a class split of 200 base categories (frequent objects) and 100 novel categories (rare objects). Following [34], we leverage image-labels obtained from MS-COCO and Open Images to learn the novel category information. We also experiment using image-caption datasets to show our method's effectiveness irrespective of training vocabulary.

Evaluation Metrics and Protocols. Following open-vocabulary methods [17, 52], for both detection and segmentation tasks, we report the mean Average Precision at intersection-over-union (IoU) of 0.5 (mAP_{50}). Following zero-shot settings [46], we report novel category performance for both *constrained setting* and *generalized setting*. In *constrained setting*, the model is evaluated only on novel class test images and in *generalized setting*, the model is evaluated on both base and novel class test images.

Implementation Details. In pseudo-mask generation framework, we use pre-trained ALBEF [25] as our vision-language model. We conducted all our pseudo-mask generation experiments using ALBEF due to the good region-text alignment when image and caption pair are present [25]. Following ALBEF, the cross-attention layer m used for

Table 1. Object Detection (mAP) performances for MS-COCO under constrained and generalized setting. C_B and C_N are subset of C_Ω, where C_Ω contains training vocabulary larger than COCO categories.

Method	Proposal Generator	Language Supervision	Base Annotation	Constrained Novel	Generalized Novel
WSDDN [5]	-	Image-labels in $C_B \cup C_N$	✗	-	19.7
Cap2Det [45]	-	Image-labels in $C_B \cup C_N$	✗	-	20.3
SB [2]	RPN $COCO_{base}$	-	✓	0.70	0.31
DELO [56]	RPN $COCO_{base}$	-	✓	7.60	3.41
PL [35]	RPN $COCO_{base}$	-	✓	10.0	4.12
OV-RCNN [46]	RPN $COCO_{base}$	Image-caption in $C_B \cup C_N$	✓	27.5	22.8
CLIP-RPN [14]	RPN $COCO_{base}$	CLIP image-text pair C_Ω	✓	-	26.3
ViLD [14]	RPN $COCO_{base}$	CLIP image-text pair C_Ω	✓	-	27.6
Detic [54]	RPN $COCO_{base}$	Image-caption in $C_B \cup C_N$	✓	-	27.8
RegionCLIP [52]	RPN $LVIS_{base}$	Conceptual caption C_Ω	✓	30.8	26.8
PB-OVD [11]	RCNN $COCO_{base}$	Image-caption in $C_B \cup C_N$	✓	32.3	30.7
XPM [17]	RPN $COCO_{base}$	Image-caption in $C_B \cup C_N$	✓	29.9	27.0
Mask-free OVIS (Ours)	WSPN $COCO_{base}$	Image-labels in $C_B \cup C_N$	✗	31.5	27.4
Mask-free OVIS (Ours)	WSPN $COCO_{base}$	Image-labels in $C_B \cup C_N$	✓	**35.9**	**31.5**

Table 2. Instance Segmentation (mAP) performances for MS-COCO and Open Images under constrained and generalized setting.

Method	Proposal Generator (MS-COCO/OpenImages)	Base Annotation	MS-COCO		Open Images	
			Constrained Novel	Generalized Novel	Constrained Novel	Generalized Novel
OVR+OMP [4]	-	✓	14.1	8.3	24.9	16.8
SB [2]	-	✓	20.8	16.0	24.8	17.3
BA-RPN [51]	-	✓	20.1	15.4	25.3	16.9
Soft-Teacher [43]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	14.8	9.6	25.9	17.6
Unbiased-Teacher [30]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	15.1	9.8	22.2	14.5
OV-RCNN [46]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	20.9	17.1	23.8	17.5
XPM [17]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	24.0	21.6	31.6	22.7
Mask-free OVIS (Ours)	WSPN $COCO_{base}$/WSPN $COCO_{base}$	✗	**27.4**	**25.0**	**35.9**	**25.8**

Grad-CAM visualization is set to 8. For attention score, we directly employ the original setting of ALBEF and no additional modification is performed. Note that other pre-trained vision-language models can also be integrated into our pipeline without major modifications. For the proposal generation pipeline, the WSPN network is trained using COCO base image-labels and the top K proposals candidates is set to 50. The WSPN network is trained for 40k iterations with learn rate 0.001 and weight decay 0.0001. For iterative masking, the hyper-parameter G is set to 3. In the segmentation pipeline, for each patch, the segmentation network is trained for 500 iterations with lr 0.25.

For fair-comparison [17, 46], we use Mask R-CNN with a ResNet50 backbone as our open-vocabulary instance segmentation model. During pseudo-mask training, we train Mask-RCNN on MS-COCO and OpenImages using batch size 8 on 8 A5000 GPUs for 90k iterations. Following [11, 14], we use text embeddings obtained from a pre-trained CLIP text encoder. During pseduo-mask training, the initial learning rate is set 0.01 and the background class weight is set to 0.2 to improve the recall of novel classes [46]. For base fine-tuning, the initial learning rate is set to 0.0005 and the weight decay is set to 0.0001. We run fine-tuning for 90k iterations where the learning rate is updated

by a decreasing factor of 0.1 at 60k and 80k iterations.

Object Detection: As shown in Table 1, we compare our method with previous established open-vocabulary detection methods on the MS-COCO dataset. Compared to weakly-supervised methods such as WSDDN [5] and zero-shot methods such as SB [2], DELO [56], our method outperforms them by a large margin. OV-RCNN [46], XPM [17] are OVD methods based on caption pre-training and our method trained with only pseudo-labels improves the novel category performance by 20.2% and 2.4% in generalized setting, respectively. Also, when compared to the method which leverages pre-trained vision-language models such as ViLD [14], RegionCLIP [52], PB-OVD [11], our method with just pseudo-labels produces similar performance. However, with fine-tuning on base annotations, our method outperforms ViLD [14], RegionCLIP [52], PB-OVD [11] by 13.3%, 14.9% and 2.8% in generalized setting, respectively. This is because with fine-tuning, the model learns task/domain specific information from noise-free annotations, boosting the novel category performance. Even without base annotations, our method outperforms most of the existing OVD methods supervised using base annotations. This shows the effectiveness of our method for learning quality representation for novel categories. Specif-

Table 3. Object detection and Instance segmentation ablation analysis for GradCAM, Pseudo-label and Mask-RCNN training.

Method	Language Supervision	Mask-RCNN Training	Base Annotation	Object Detection		Instance Segmentation	
				Constrained	Generalized	Constrained	Generalized
				Novel/Base	Novel/Base/All	Novel/Base	Novel/Base/All
GradCAM	$\mathcal{C}_N$	✗	✗	8.6/0.0	-	5.2/0.0	-
PL	$\mathcal{C}_N$	✗	✗	17.3/0.0	-	14.8/0.0	-
PL + Mask-RCNN	$\mathcal{C}_N$	✓	✗	31.1/0.4	27.1/0.6/7.6	27.0/0.5	24.7/0.5/6.9
PL + Mask-RCNN	$\mathcal{C}_B \cup \mathcal{C}_N$	✓	✗	32.4/22.4	29.3/22.8/24.5	27.4/18.4	25.0/18.3/20.1
PL + Mask-RCNN	$\mathcal{C}_B \cup \mathcal{C}_N$	✓	✓	35.9/40.7	40.0/31.5/37.7	31.2/36.7	36.0/28.7/34.0

ically, the quality representation is learned due to the quality proposal generated by WSPN compared to fully-supervised RPN and RCNN proposal generators.

Instance Segmentation: Table 2 compares our method with previous open-vocabulary instance segmentation methods on the MS-COCO and Open Images datasets. SB [2] and BA-RPN [51] are zero-shot methods which utilize different background modelling strategies and caption pre-training to learn novel categories. Compared to these, our method improves the novel category performance by a large margin on both datasets and settings. When compared against conventional pseudo-labelling methods, such as soft-teacher [43] and unbiased teacher [30], our method significantly improves on MS-COCO and Open Images datasets. Finally, when compared to open-vocabulary methods such as OV-RCNN [46], XPM [17], our method outperforms by 4.6 and 3.1 mAP in COCO and Open Images in generalized setting, respectively. All these comparisons are performed against our method trained with just pseudo-labels and no base annotation is used during training. This shows the effectiveness of our overall pseudo-mask generation pipeline.

4.1. Ablation Study

GradCAM vs Pseudo-Labels vs Mask-RCNN Training. In Table 3, we analyze the quality of pseudo-labels generated from the GradCAM activation map and our method (PL). After pseudo-label generation, we show how Mask-RCNN training helps to improve the quality of prediction compared to pseudo-labels. Finally, we show how fine-tuning on base annotation improves our method. In the first row, we evaluate the pseudo-mask for novel samples, where the pseudo-mask is generated by normalizing and threshold the GradCAM activation map. In the second row, we evaluate the pseudo-mask for novel samples, where the pseudo-mask is generated by our method. From Table 3, we can observe the quality of the pseudo-mask generated by our method for novel samples is much better than the GradCAM activation map as a pseudo-mask. From Table 3 third row, we can observe training a Mask-RCNN on pseudo-labels improves the performance on novel categories by modelling fine-grained information. By including pseudo-labels from base categories, we observe that the performance on novel samples further improves. Finally, when fine-tuning on base annotations, the performance on novel categories signifi-

Table 4. Ablation analysis for object detection and instance segmentation under different language supervision.

Method	Language Supervision	Object Detection		Instance Segmentatn	
		Constrnd Novel	Genrlzd Novel	Constrnd Novel	Genrlzd Novel
OV-RCNN [46]	Image-caption COCO	27.5	22.8	-	-
PB-OVD [11]	Image-caption COCO	-	29.1	-	-
PB-OVD [11]	Img-Cap COCO, SBU, VG	32.3	30.8	-	-
Ours	Image-labels COCO	35.9	31.5	31.0	28.3
Ours	Image-caption COCO	36.1	31.8	31.5	28.8

Table 5. Ablation analysis for object detection under different proposal generator. All models are fined-tuned on COCO base.

Method	Proposal Generator	Constrained Novel	Generalized Novel
OV-RCNN [46]	RPN COCO Base	27.8	22.8
PB-OVD [11]	RCNN COCO Base	32.3	30.8
Ours	Selective Search	34.5	31.0
Ours	WSPN COCO Base	35.9	31.5

cantly improves by learning task/domain specific information from noise-free manual annotations.

Language supervision. Given an image and caption pair, our method can generate a pseudo-mask leveraging a pre-trained vision-language model. Thus to analyze the effect of captions, we conduct experiments between human-provided captions and pseudo-captions generated from image-labels. As show in Table 4, human-provided captions and image-labels based pseudo-caption produce similar performance showing that irrespective of caption type, our method can generate pseudo-mask for the object of interest (see supplementary material for visual comparison). Therefore, our method is more data efficient as it requires cheap image-labels compared to human-provided captions. Table 4 compares our method with other pseudo-label generation methods trained with extra language supervision [11]. Our method, with lesser language supervision, outperforms [11] by a considerable margin.

Proposal Generator Quality vs Performance. In general, better quality of proposal provides better quality of pseudo-labels. Therefore, we generated pseduo-labels using different proposal generator and the results are reported in Table 5. As shown in Table 5, our method trained with WSPN as proposal generator is produces better performance compared to methods which rely on fully-supervised proposal generator such as RPN and RCNN. Also when compared to selective search as proposal generator, WSPN demonstrates better performance for novel categories. This is because WSPN refines selective search proposals and localizes them
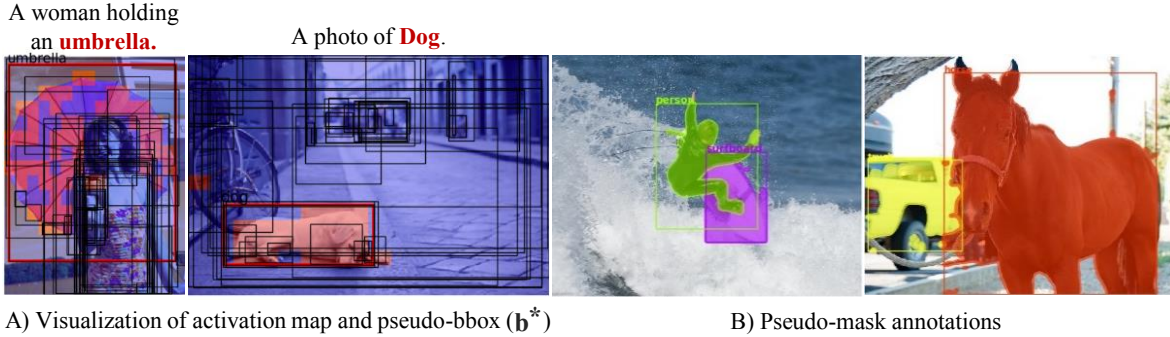
A woman holding an **umbrella.**

A photo of **Dog**.

A) Visualization of activation map and pseudo-bbox ($\mathbf{b}^*$)

B) Pseudo-mask annotations

Figure 5. A) Pseduo bounding box selection guided by GradCAM activation. B) Visualization of pseudo-mask generated for Open Images.

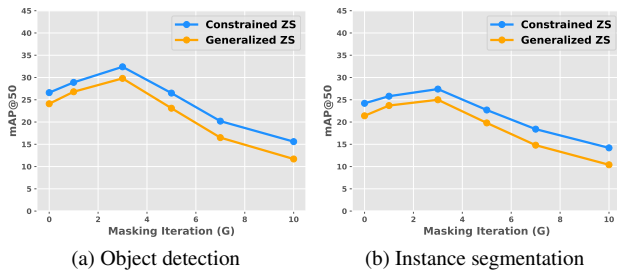(a) Object detection

(b) Instance segmentation

Figure 6. Ablation study for object detection and instance segmentation by increase number of masking iteration during PL generation. As iteration increases, GradCAM activation becomes noisy resulting in poor guidance and pseudo-mask generation.

towards objects producing better quality proposals.

Iterative Masking Steps vs Performance. Fig. 6 presents ablation analysis for iterative masking hyper-parametr G. We see that for initial few masking steps the performance increase. Because, without any masking the GradCAM activation map is towards most discriminative parts of an object. As we mask out these region, the less discriminative regions gets activated and combining activation from previous steps produce a strong guidance function Φ_t. However, performing too many mask iterations could completely mask out the object of interest and unrelated background regions might get activated resulting in a poor Φ_t.

Weakly-supervised Instance Segmentation. Weakly-supervised Instance Segmentation (WSIS) methods performs instance-segmentation using supervision from image-labels. Similar to WSIS setting, our method performs instance-segmentation using just image-labels. Following this, we simply extended our method to WSIS setting by comparing with VOC benchmark. PENet [12] is a recent WSIS method, which relies on multiple downstream task network such as segmentation, detection and multi-class classification to perfrom segmentation. CL [18] and IAM [57] are WSIS methods utilize advanced attention mechanism to generate instance segmentation mask. As shown in Table 6, our method outperform existing WSIS methods by considerable margin. Thus, the proposed method is simple and can be easily extended to OV setting or WSIS setting producing SOTA performance. Note that, other OV methods cannot be extended to WSIS setting as they rely on base

Table 6. Instance segmentation(mAP) performance for VOC in weakly-supervised instance segmentation setting.

Method	Supervision	mAP
Mask-RCNN [15]	F	67.9
OCIS [8]	I	26.8
PRM [55]	I	28.3
IAM [57]	I	30.2
CL [18]	I+C	30.2
PENet [12]	I	38.1
Ours	I	**38.6**

annotation and similarly WSIS methods cannot be extended to OV setting due to constrained vocabulary space.

Qualitative Analysis. Fig. 5 (A) presents a visualization of activation maps generated for the object of interest (woman and dog). As we can see, the generated activation map covers the entire object and it can be used as a guidance function to choose the best bounding box proposal. Note that the activation are square-shaped because the original activation map is 1/16'th of the image size. We perform nearest interpolation to obtain an activation map of image size. In Fig 5 (B), we visualize the pseudo-mask for Open Images generated from our pipeline. We can observe that the generated pseudo-mask is of good quality; still contains some false positives. However, with Mask-RCNN training, the model learns to filter the noise present in the pseudo-mask producing better-quality predictions.

5. Conclusion

We propose a novel pipeline called Mask-free OVIS for open-vocabulary instance segmentation, which eliminates the need for costly and time-consuming manual mask annotations. We achieve this by generating pseudo-mask annotation for base and novel categories by leveraging a pretrained vision-language model and weakly supervised techniques. The generated pseudo-mask annotations are then used to supervise the Mask-RCNN model. Extensive experiments are conducted to show the effectiveness of our method on MS-COCO and Open Images datasets. Instead of relying on labour-expensive instance-level annotations, we hope our simple and effective pipeline provides an alternative to detect and segment long-tailed novel categories.

References

[1] Aditya Arun, CV Jawahar, and M Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9432–9441, 2019. 3

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 6, 7

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2824–2833, 2021. 6

[5] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2846–2854, 2016. 4, 6

[6] Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 393–408. Springer, 2022. 3

[7] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4013–4022, 2018. 1

[8] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12397–12405, 2019. 8

[9] 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. 3

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[11] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 266–282. Springer, 2022. 1, 3, 6, 7

[12] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3345–3354, 2019. 8

[13] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[14] 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. 1, 2, 3, 6

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 5, 8

[16] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in neural information processing systems*, 33:16797–16807, 2020. 3

[17] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2, 3, 5, 6, 7

[18] Jaedong Hwang, Seohyun Kim, Jeany Son, and Bohyung Han. Weakly supervised instance segmentation by deep community learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1020–1029, 2021. 8

[19] 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

[20] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing*, 29:8055–8068, 2020. 2

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 5

[22] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 3

[23] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 3

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for uni-

fied vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3, 5

[26] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

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

[29] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[30] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations*, 2021. 6, 7

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[32] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 3

[33] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Science and information conference*, pages 128–144. Springer, 2019. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[35] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020. 6

[36] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022. 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 4

[38] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020. 4

[39] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 3, 4

[40] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 3

[41] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 3

[42] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020. 1

[43] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 6, 7

[44] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8381, 2019. 3

[45] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9686–9695, 2019. 6

[46] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2, 3, 5, 6, 7

[47] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection.

In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8292–8300, 2019. 3

[48] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021. 3

[49] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[50] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 159–175. Springer, 2022. 2, 3

[51] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2593–2602, 2021. 6, 7

[52] 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. 1, 2, 3, 5, 6

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[54] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 6

[55] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3791–3800, 2018. 8

[56] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020. 6

[57] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3116–3125, 2019. 8