Exploiting Unlabeled Data with Vision and Language Models for Object Detection

Shiyu Zhao[1,⋆], Zhixing Zhang[1,⋆], Samuel Schulter[2], Long Zhao[3], Vijay Kumar B.G[2], Anastasis Stathopoulos[1], Manmohan Chandraker[2,4], and Dimitris Metaxas[1]

[1]Rutgers University, [2]NEC Labs America, [3]Google Research, [4]UC San Diego

Abstract. Building robust and generic object detection frameworks requires scaling to larger label spaces and bigger training datasets. However, it is prohibitively costly to acquire annotations for thousands of categories at a large scale. We propose a novel method that leverages the rich semantics available in recent vision and language models to localize and classify objects in unlabeled images, effectively generating pseudo labels for object detection. Starting with a generic and class-agnostic region proposal mechanism, we use vision and language models to categorize each region of an image into any object category that is required for downstream tasks. We demonstrate the value of the generated pseudo labels in two specific tasks, open-vocabulary detection, where a model needs to generalize to unseen object categories, and semi-supervised object detection, where additional unlabeled images can be used to improve the model. Our empirical evaluation shows the effectiveness of the pseudo labels in both tasks, where we outperform competitive baselines and achieve a novel state-of-the-art for open-vocabulary object detection. Our code is available at `https://github.com/xiaofeng94/VL-PLM`.

1 Introduction

Recent advances in object detection build on large-scale datasets [17,27,41], which provide rich and accurate human-annotated bounding boxes for many object categories. However, the annotation cost of such datasets is significant. Moreover, the long-tailed distribution of natural object categories makes it even harder to collect sufficient annotations for all categories. Semi-supervised object detection (SSOD) [44,60] and open-vocabulary object detection (OVD) [4,16,54] are two tasks to lower annotations costs by leveraging different forms of unlabeled data. In SSOD, a small fraction of fully-annotated training images is given along with a large corpus of unlabeled images. In OVD, a fraction of the desired object categories is annotated (the base categories) in all training images and the task is to also detect a set of novel (or unknown) categories at test time. These object categories can be present in the training images, but are not annotated with ground truth bounding boxes. A common and successful approach for leveraging

⋆ Equal contribution.

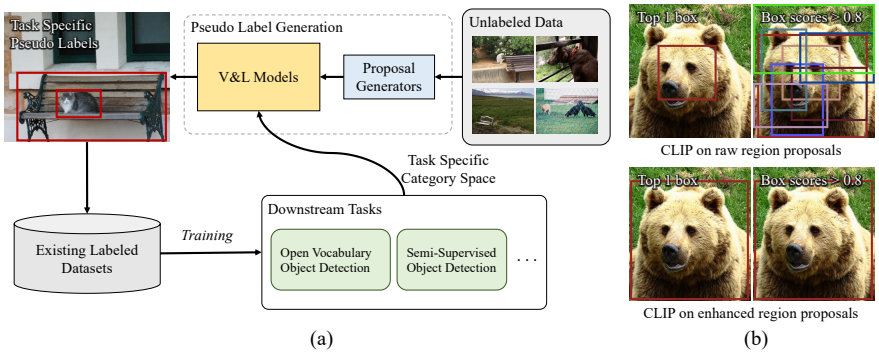

Fig. 1. (a) Overview of leveraging the semantic knowledge contained in vision and language models for mining unlabeled data to improve object detection systems for open-vocabulary and semi-supervised tasks. **(b)** Illustration of the weak localization ability when applying CLIP [37] on raw object proposals (top), compared with our improvements (bottom). The left images show the pseudo label with the highest score. The right images show all pseudo labels with scores greater than 0.8. The proposed scoring gives much cleaner pseudo labels.

unlabeled data is by generating pseudo labels. However, all prior works on SSOD only leveraged the small set of labeled data for generating pseudo labels, while most prior work on OVD does not leverage pseudo labels at all.

In this work, we propose a simple but effective way to mine unlabeled images using recently proposed vision and language (V&L) models to generate pseudo labels for both known and unknown categories, which suits both tasks, SSOD and OVD. V&L models [23,29,37] can be trained from (noisy) image caption pairs, which can be obtained at a large scale without human annotation efforts by crawling websites for images and their alt-texts. Despite the noisy annotations, these models demonstrate excellent performance on various semantic tasks like zero-shot classification or image-text retrieval. The large amount of diverse images, combined with the free-form text, provides a powerful source of information to train robust and generic models. These properties make vision and language models an ideal candidate to improve existing object detection pipelines that leverage unlabeled data, like OVD or SSOD, see Fig. 1(a).

Specifically, our approach leverages the recently proposed vision and language model CLIP [37] to generate pseudo labels for object detection. We first predict region proposals with a two-stage class-agnostic proposal generator which was trained with limited ground truth (using only known base categories in OVD and only labeled images in SSOD), but generalizes to unseen categories. For each region proposal, we then obtain a probability distribution over the desired object categories (depending on the task) with the pre-trained V&L model CLIP [37]. However, as shown in Fig. 1(b), a major challenge of V&L models is the rather low object localization quality, also observed in [57]. To improve localization, we propose two strategies where the two-stage proposal generator helps the V&L

model: (1) Fusing CLIP scores and objectness scores of the two-stage proposal generator, and (2) removing redundant proposals by repeated application of the localization head (2nd stage) in the proposal generator. Finally, the generated pseudo labels are combined with the original ground truth to train the final detector. We name our method as **V**&**L**-guided **P**seudo-**L**abel **M**ining (VL-PLM).

Extensive experiments demonstrate that VL-PLM successfully exploits the unlabeled data for open-vocabulary detection and outperforms the state-of-the-art ViLD [16] on novel categories by +6.8 AP on the COCO dataset [32]. Moreover, VL-PLM improves the performance on known categories in SSOD and beats the popular baseline STAC [44] by a clear margin, by only replacing its pseudo labels with ours. Besides, we also conduct various ablation studies on the properties of the generated pseudo labels and analyze the design choices of our proposed method. We also believe that VL-PLM can be further improved with better V&L models like ALIGN [23] or ALBEF [29].

The contributions of our work are as follows: **(1)** We leverage V&L models for improving object detection frameworks by generating pseudo labels on unlabeled data. **(2)** A simple but effective strategy to improve the localization quality of pseudo labels scored with the V&L model CLIP [37]. **(3)** State-of-the-art results for novel categories on the COCO open-vocabulary detection setting. **(4)** We showcase the benefits of VL-PLM in a semi-supervised object detection setting.

2 Related Work

The goal of our work is to improve object detection systems by leveraging unlabeled data via vision and language models that carry rich semantic information.

Vision & language (VL) models: Combining natural language and images has enabled many valuable applications in recent years, like image captioning [2,7,12,25], visual question answering [1,13,20,30,36,55], referring expression comprehension [8,24,26,34,35,52,53], image-text retrieval [29,37,47] or language-driven embodied AI [3,9]. While early works proposed task-specific models, generic representation learning from vision and language inputs has gained more attention [8,19,33,34,45]. Most recent works like CLIP [37] or ALIGN [23] also propose generic vision and language representation learning approaches, but have significantly increased the scale of training data, which led to impressive results in tasks like zero-shot image classification or image-text retrieval. The training data consist of image and text pairs, typically crawled from the web at a very large scale (400M for [37] and 1.2B for [23]), but without human annotation effort. In our work, we leverage such pre-trained models to mine unlabeled data and to generate pseudo labels in the form of bounding boxes, suitable for object detection. One challenge with using such V&L models [23,37] is their limited capability in localizing objects (recall Fig. 1(b)), likely due to the lack of region-word alignment in the image-text pairs of their training data. In Sec. 3.2, we show how to improve localization quality with our proposal generator.

Vision & language models for dense prediction tasks: The success of CLIP [37] (and others [23,29]) has motivated the extension of zero-shot classifica-

tion capabilities to dense image prediction tasks like object detection [16,21,42,54] or semantic segmentation [28,39,50,59]. These works try to map features of individual objects (detection) or pixels (segmentation) into the joint vision-language embedding space provided by models like CLIP. For example, ViLD [16] trains an object detector in the open-vocabulary regime by predicting the text embedding (from the CLIP text-encoder) of the category name for each image region. LSeg [28] follows a similar approach, but is applied to zero-shot semantic segmentation. Both works leverage task-specific insights and do not generate explicit pseudo labels. In contrast, our proposed VL-PLM is more generic by generating pseudo labels, thus enabling also other tasks like semi-supervised object detection [44]. Similar to our work, both Gao *et al.* [14] and Zhong *et al.* [57] generate explicit pseudo labels in the form of bounding boxes. In [14], the attention maps of a pretrained V&L model [29] between words of a given caption and image regions are used together with object proposals to generate pseudo labels. In contrast, our approach does not require image captions as input and we use only unlabeled images, while still outperforming [14] in an open-vocabulary setting on COCO. RegionCLIP [57] assigns semantics to region proposals via a pre-trained V&L model, effectively creating pseudo labels in the form of bounding boxes. While our approach uses such pseudo labels directly for training object detectors, [57] uses them for fine-tuning the original V&L model, which then builds the basis for downstream tasks like open-vocabulary detection. We believe this contribution is orthogonal to ours as it effectively builds a better starting point of the V&L model, and can be incorporated into our framework as well. Interestingly, even without the refined V&L model, we show improved accuracy with pseudo labels specifically for novel categories as shown in Sec. 4.1.

The main focus of all the aforementioned works is to enable the dynamic expansion of the label space and to recognize novel categories. While our work also demonstrates state-of-the-art results in this open-vocabulary setting, where we mine unlabeled data for novel categories, we want to stress that our pseudo labels are applicable more generally. In particular, we also use a V&L model to mine unlabeled images for known categories in a semi-supervised object detection setting. Furthermore, by building on the general concept of pseudo labels, our approach may be extended to other dense prediction tasks like semantic segmentation in future works as well.

Object detection from incomplete annotations: Pseudo labels are proven useful in many recent object detection methods trained with various forms of weak annotations: semi-supervised detection [44,60], unsupervised object discovery [43], open-vocabulary detection [14,57], weakly-supervised detection [10,58], unsupervised domain adaptation [22,51] or multi-dataset detection [56]. In all cases, an initial model trained from base information is applied on the training data to obtain the missing information. Our main proposal is to leverage V&L models to improve these pseudo labels and have one unified way of improving the accuracy in multiple settings, see Sec. 3.3. In this work, we focus on two important forms of weak supervision: zero-shot/open-vocabulary detection (OVD) and semi-supervised object detection (SSOD). In zero-shot detection [4] a model

is trained from a set of base categories. Without ever seeing any instance of a novel category during training, the model is asked to predict novel categories, typically via association in a different embedding space, like attribute or text embeddings. Recent works [16,38,54] relax the setting to include novel categories in the training data, but without bounding box annotations, which also enables V&L models to be used (via additional images that come with caption data). ViLD [16], as described above, uses CLIP [37] with model distillation losses to make predictions in the joint vision-text embedding space. In contrast, we demonstrate that explicitly creating pseudo labels for novel categories via mining the training data can significantly improve the accuracy, see Sec. 4.1. The second task we focus on is semi-supervised object detection (SSOD), where a small set of images with bounding box annotations and a large set of unlabeled images are given. In contrast to OVD, the label space does not change from train to test time. A popular and recent baseline that builds on pseudo labels is STAC [44]. This approach employs a consistency loss between predictions on a strongly augmented image and pseudo labels computed on the original image. We demonstrate the benefit of leveraging V&L models to improve the pseudo label quality in such a framework. Other works on SSOD, like [49,60] propose several orthogonal improvements which can be incorporated into our framework as well. In this work, however, we focus purely on the impact of the pseudo labels. Finally, note that our concepts may also be applicable to other tasks beyond open-vocabulary and semi-supervised object detection, but we leave this for future work.

3 Method

The goal of our work is to mine unlabeled images with vision & language (V&L) models to generate semantically rich pseudo labels (PLs) in the form of bounding boxes so that object detectors can better leverage unlabeled data. We start with a generic training strategy for object detectors with the unlabeled data in Sec. 3.1. Then, Sec. 3.2 describes the proposed VL-PLM for pseudo label generation. Finally, Sec. 3.3 presents specific object detection tasks with our PLs.

3.1 Training object detectors with unlabeled data

Unlabeled data comes in many different forms for object detectors. In semi-supervised object detection, we have a set of fully-labeled images $\mathcal{I}_L$ with annotations for the full label space $\mathcal{S}$, as well as unlabeled images $\mathcal{I}_U$, with $\mathcal{I}_L \cap \mathcal{I}_U = \varnothing$. In open-vocabulary detection, we have partly-labeled images with annotations for the set of base categories $\mathcal{S}_B$, but without annotations for the unknown/novel categories $\mathcal{S}_N$. Note that partly-labeled images are therefore contained in both $\mathcal{I}_L$ and $\mathcal{I}_U$, i.e., $\mathcal{I}_L = \mathcal{I}_U$.

A popular and successful approach to learn from unlabeled data is via pseudo labels. Recent semi-supervised object detection methods follow this approach by first training a teacher model on the limited ground truth data, then generating pseudo labels for the unlabeled data, and finally training a student model. In the

following, we describe a general training strategy for object detection to handle different forms of unlabeled data.

We define a generic loss function for an object detector with parameters θ over both labeled and unlabeled images as

$$\mathcal{L}(\theta, \mathcal{I}) = \frac{1}{N_{\mathcal{I}}} \sum_{i=1}^{N_{\mathcal{I}}} [I_i \in \mathcal{I}_L] \; l_s(\theta, I_i) + \alpha [I_i \in \mathcal{I}_U] \; l_u(\theta, I_i) \; , \tag{1}$$

where α is a hyperparameter to balance supervised l_s and unsupervised l_u losses and $[\cdot]$ is the indicator function returning either 0 or 1 depending on the condition. Note again that I_i can be contained in both $\mathcal{I}_L$ and $\mathcal{I}_U$.

Object detection ultimately is a set prediction problem and to define a loss function, the set of predictions (class probabilities and bounding box estimates) need to be matched with the set of ground truth boxes. Different options exist to find a matching [6,18] but it is mainly defined by the similarity (IoU) between predicted and ground truth boxes. We define the matching for prediction i as $\sigma(i)$, which returns a ground truth index j if successfully matched or `nil` otherwise. The supervised loss l_s contains a standard cross-entropy loss for the classification l_{cls} and an ℓ_1 loss for the box regression l_{reg}. Given $I \in \mathcal{I}$, we define l_s as,

$$l_s(\theta, I) = \frac{1}{N^*} \sum_i l_{cls} \left(C_i^\theta(I), c^*_{\sigma(i)} \right) + [\sigma(i) \neq \texttt{nil}] \; l_{reg} \left(T_i^\theta(I), \mathbf{t}^*_{\sigma(i)} \right) \; , \tag{2}$$

where N^* is the number of predicted bounding boxes. $C_i^\theta(\cdot)$ and $T_i^\theta(\cdot)$ are the predicted class distributions and bounding boxes of the object detector. The corresponding (matched) ground truth is defined as $c^*_{\sigma(i)}$ and $\mathbf{t}^*_{\sigma(i)}$, respectively.

The unsupervised loss l_u is similarly defined, but uses pseudo labels with high confidence as supervision signals:

$$\begin{aligned} l_u(\theta, I) = \frac{1}{N^u} \sum_i [\max(\mathbf{p}^u_{\sigma(i)}) \geq \tau] \cdot \Big(& l_{cls} \left(C_i^\theta(I), \hat{c}^u_{\sigma(i)} \right) + \\ & [\sigma(i) \neq \texttt{nil}] \; l_{reg} \left(T_i^\theta(I), \mathbf{t}^u_{\sigma(i)} \right) \Big) \; . \end{aligned} \tag{3}$$

Here, $\mathbf{p}^u_{\sigma(i)}$ defines the probability distribution over the label space of the pseudo label matched with prediction i and N^u is the number of adopted pseudo labels, i.e., $N^u = \sum_i [\max(\mathbf{p}^u_{\sigma(i)}) \geq \tau]$. Pseudo labels for the classification and the box regression losses are $\hat{c}^u_{\sigma(i)} = \arg\max(\mathbf{p}^u_{\sigma(i)})$ and $\mathbf{t}^u_{\sigma(i)}$, respectively.

The key to successful training of object detectors from unlabeled data are accurate pseudo labels. In the next section, we will present our approach, VL-PLM, to leverage V&L models as external models to exploit unlabeled data for generating pseudo labels.

3.2 VL-PLM: Pseudo labels from vision & language models

V&L models are trained on large scale datasets with image-text pairs that cover a diverse set of image domains and rich semantics in natural text. Moreover, the

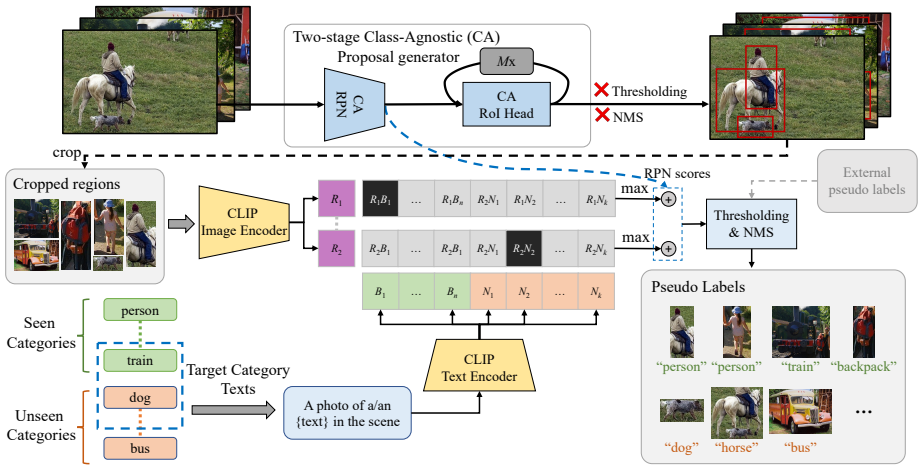

Fig. 2. Overview of the proposed VL-PLM to mine unlabeled images with vision & language models to generate pseudo labels for object detection. The top part illustrates our class-agnostic proposal generator, which improves the pseudo label localization by using the class-agnostic proposal score and the repeated application of the RoI head. The bottom part illustrates the scoring of cropped regions with the V&L model based on the target category names. The chosen category names can be adjusted for the desired downstream task. After thresholding and NMS, we get the final pseudo labels. For some tasks like SSOD, we will merge external pseudo labels for a teacher model with ours before thresholding and NMS.

image-text pairs can be obtained without costly human annotation by using web-crawled data (images and corresponding alt-texts) [37,23]. Thus, V&L models are ideal sources of external knowledge to generate pseudo labels for arbitrary categories, which can be used for downstream tasks like open-vocabulary or semi-supervised object detection.

Overview: Fig. 2 illustrates the overall pipeline of our pseudo label generation with the recent V&L model CLIP [37]. We first feed an unlabeled image into our two-stage class-agnostic detector (described in the next section below) to obtain region proposals. We then crop image patches based on those regions and feed them into the CLIP image-encoder to obtain an embedding in the CLIP vision-and-language space. Using the corresponding CLIP text-encoder and template text prompts, we generate embeddings for category names that are desired for the specific task. For each region, we compute the similarities between the region embedding and the text embeddings via a dot product and use softmax to obtain a distribution over the categories. We then generate the final pseudo labels using scores from both class-agnostic detector and V&L model, which we describe in detail below.

There are two key challenges in our framework: (1) Generating robust proposals for novel categories, required by open-vocabulary detection, and (2) overcoming

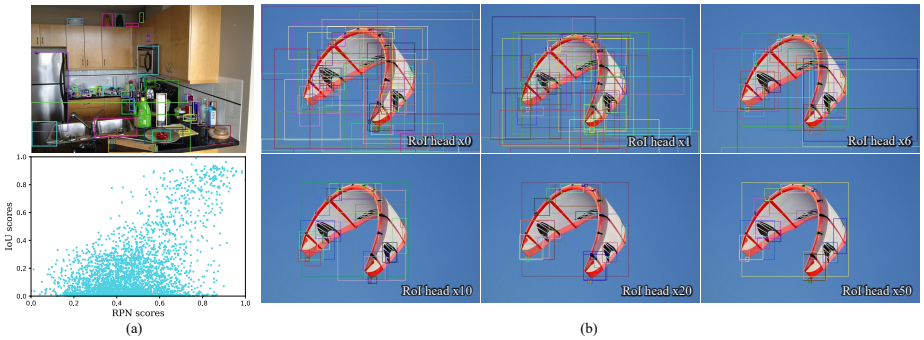

Fig. 3. (a) RPN scores indicate localization quality. Top: Top 50 boxes from RPN in an image which correctly locates nearly all objects. Bottom: A positive correlation between RPN and IoU scores for RPN boxes of 50 randomly sampled COCO images. The correlation coefficient is 0.51. **(b)** Box refinement by repeating RoI head. "×N" indicates how many times we repeat the RoI head.

the poor localization quality of the raw CLIP model, see Fig. 1(b). We introduce simple but effective solutions to address the two challenges in the following.

Generating robust and class-agnostic region proposals: To benefit tasks like open vocabulary detection with the unlabeled data, the proposal generator should be able to locate not only objects of categories seen during training but also of objects of novel categories. While unsupervised candidates like selective search [46] exist, these are often time-consuming and generate many noisy boxes. As suggested in prior studies [16,54], the region proposal network (RPN) of a two-stage detector generalizes well for novel categories. Moreover, we find that the RoI head is able to improve the localization of region proposals, which is elaborated in the next section. Thus, we train a standard two-stage detector, e.g., Faster-RCNN [40], as our proposal generator using available ground truth, which are annotations of base categories for open vocabulary detection and annotations from the small fraction of annotated images in semi-supervised detection. To further improve the generalization ability, we ignore the category information of the training set and train a class-agnostic proposal generator. Please refer to Sec. 4.3 and the supplement for a detailed analysis of the proposal generator.

Generating pseudo labels with a V&L model: Directly applying CLIP [37] on cropped region proposals yields low localization quality, as was observed in Fig. 1(b) and also in [57]. Here, we demonstrate how to improve the localization ability with our two-stage class-agnostic proposal generator in two ways. Firstly, we find that the RPN score is a good indicator for localization quality of region proposals. Fig. 3(a) illustrates a positive correlation between RPN and IoU scores. We leverage this observation and average the RPN score with those of the CLIP predictions. Secondly, we remove thresholding and NMS of the proposal generator and feed proposal boxes into the RoI head multiple times, similar to [5]. We observe that it pushes redundant boxes closer to each other by repeating the RoI

head, which can be seen in Fig. 3(b). In this way, we encounter better located bounding boxes and provide better pseudo labels. Please refer to Sec. 4.3 for a corresponding empirical analysis.

To further improve the quality of our pseudo labels, we adopt the multi-scale region embedding from CLIP as described in [16]. Moreover, as suggested in [44], we employ a high threshold to pick pseudo labels with high confidence. The confidence score of the pseudo label for the region R_i is formulated as $\overline{c}_i^u = [s_i^u \geq \tau] \cdot s_i^u$, with

$$s_i^u = \frac{S_{RPN}(R_i) + \max(\mathbf{p}_i^u)}{2} \,, \tag{4}$$

where $S_{RPN}(\cdot)$ denotes the RPN score. The prediction probability distribution $\mathbf{p}_i^u$ is defined as

$$\mathbf{p}_i^u = \text{softmax}\{\phi(E_{\text{im}}(R_i) + E_{\text{im}}(R_i^{1.5\times})) \cdot E_{\text{txt}}(\text{Categories})^T\}. \tag{5}$$

Here, $R_i^{1.5\times}$ is a region cropped by $1.5\times$ the size of R_i. E_{im} and E_{txt} are the image and text encoders of CLIP, respectively, and $\phi(\mathbf{x}) = \mathbf{x}/||\mathbf{x}||$. If $\overline{c}_i^u = 0$, we exclude R_i from our pseudo labels.

3.3 Using our pseudo labels for downstream tasks

Finally, we briefly describe how we use the pseudo labels that are generated from unlabeled data for two specific downstream tasks that we focus on in this work.
Open-vocabulary detection: In this task, the detector has access to images with annotations for base categories and needs to generalize to novel categories. We leverage the data of the base categories to train a class-agnostic Mask R-CNN as our proposal generator and take the names of novel categories as the input texts of the CLIP text-encoder in aforementioned pseudo label generation process. Then, we train a standard Mask R-CNN with RestNet50-FPN [31] with both base ground truth and novel pseudo labels as described in Sec. 3.1.
Semi-supervised object detection: In this task, relevant methods usually train a teacher model using ground truth from the limited set of labeled images, and then generate pseudo labels with the teacher on the unlabeled images. We also generate those pseudo labels and merge them with pseudo labels from our VL-PLM. Please refer to the supplementary document for details. Thus, the student model is trained on available ground truth and pseudo labels from both our V&L-based approach and the teacher model.

4 Experiments

We experimentally evaluate the proposed VL-PLM first on open-vocabulary detection in Sec. 4.1 and then on semi-supervised object detection in Sec. 4.2. In Sec. 4.3 we ablate various design choices of VL-PLM.

Table 1. Evaluations for open vocabulary detection on the COCO 2017 [32]. RegionCLIP* indicates a model without refinement using image-caption pairs.

Method	Training Source	Novel AP	Base AP	Overall AP
Bansal *et al.* [4]	instance-level labels in $\mathcal{S}_B$	0.31	29.2	24.9
Zhu *et al.* [61]		3.41	13.8	13.0
Rahman *et al.* [38]		4.12	35.9	27.9
OVR-CNN [54]	image-caption pairs in $\mathcal{S}_B \cup \mathcal{S}_N$ instance-level labels in $\mathcal{S}_B$	22.8	46.0	39.9
Gao *et al.* [14]	raw image-text pairs via Internet image-caption pairs in $\mathcal{S}_B \cup \mathcal{S}_N$ instance-level labels in $\mathcal{S}_B$	30.8	46.1	42.1
RegionCLIP [57]		31.4	57.1	50.4
RegionCLIP* [57]	raw image-text pairs via Internet instance-level labels in $\mathcal{S}_B$	14.2	52.8	42.7
ViLD [16]		27.6	59.5	51.3
VL-PLM (Ours)		**34.4**	**60.2**	**53.5**

4.1 Open-vocabulary object detection

In this task, we have a training set with annotations for known base categories $\mathcal{S}_B$. Our goal is to train a detector for novel categories $\mathcal{S}_N$. Usually, the labeled images $\mathcal{I}_L$ and the unlabeled images $\mathcal{I}_U$ are the same, i.e., $\mathcal{I}_L = \mathcal{I}_U$.

Experimental setup: Following prior studies [4,14,16,54], we base our evaluation on COCO 2017 [32] in the zero-shot setting (COCO-ZS) where there are 48 known base categories and 17 unknown novel categories. Images from the training set are regarded as labeled for base classes and also as unlabeled for novel classes. We take the widely adopted mean Average Precision at an IoU of 0.5 (AP_{50}) as the metric and mainly compare our method with ViLD [16], the state-of-the-art method for open vocabulary detection. Thus, we follow ViLD and report AP_{50} over novel categories, base categories and all categories as Novel AP, Base AP, and Overall AP, respectively. Our supplemental material contains results for the LVIS [17] dataset.

Implementation details: We set a NMS threshold of 0.3 for the RPN of the proposal generator. The confidence threshold for pseudo labels (PLs) is $\tau = 0.8$. Finally, we obtain an average of 4.09 PLs per image, which achieve a Novel AP of 20.9. We use the above hyperparameters for pseudo label generation in all experiments, unless otherwise specified. The proposal generator and the final detector were implemented in Detectron2 [48] and trained on a server with NVIDIA A100 GPUs. The proposal generator was trained for 90,000 iterations with a batch size of 16. Similar to ViLD, the final detector is trained from scratch for 180,000 iterations with input size of 1024×1024, large-scale jitter augmentation [15], synchronized batch normalization of batch size 128, weight decay of 4e-5, and an initial learning rate of 0.32.

Comparison to SOTA: As shown in Table 1, the detector trained with VL-PLM significantly outperforms the prior state-of-the-art ViLD by nearly +7% in Novel AP. Compared with [54] and [14], our method achieves much better

Table 2. Open-vocabulary models trained with base categories from COCO are evaluated on unseen datasets. The evaluation protocol follows [14] and reports AP50

PLs	Iterations×Batch size	VOC 2007	Object365	LVIS
Gao *et al.* [14]	150K×64	59.2	6.9	8.0
VL-PLM	180K×16	**67.4**	**10.9**	**22.2**

performance not only on novel but also on base categories. This indicates training with our PLs has less impact on the predictions of base categories, where previous approaches suffered a huge performance drop. Overall, we can see that using V&L models to explicitly generate PLs for novel categories to train the model can give a clear performance boost. Although this introduces an overhead compared to ViLD (and others), which can include novel categories dynamically into the label space, many practical applications easily tolerate this overhead in favor of significantly improved accuracy. Such a setup is also similar to prior works that generate synthetic features of novel categories [61]. Moreover, our method has large potential for further improvement with better V&L model. [16] demonstrates a 60% performance boost of ViLD when using ALIGN [23] as the V&L model. We expect similar improvements on VL-PLM if ALIGN is available.

Generalizing to unseen datasets: Following Gao *et al.*'s evaluation protocol [14], we evaluate COCO-trained models on three unseen datasets: VOC 2007 [11], Object365 [41] and LVIS [17]. To do so, we generate PLs for the novel label spaces of these datasets on the COCO dataset and train a standard Faster R-CNN model. The results of our approach on the three unseen datasets is compared to [14] in Table 2. VL-PLM significantly outperforms [14] with similar iterations and smaller batch sizes. Note that [14] requires additional image captions to generate PLs, while VL-PLM can generate PLs for any given category.

4.2 Semi-supervised object detection

In this task, we have annotations for all categories on a small portion of a large image set. This portion is regarded as the labeled set $\mathcal{I}_L$ and the remaining images are regarded as the unlabeled set $\mathcal{I}_U$ i.e. $\mathcal{I}_L \cap \mathcal{I}_U = \varnothing$.

Experimental setup: Following previous studies [44,49,60], we conduct experiments on COCO [32] with 1, 2, 5, and 10% of the training images selected as the labeled data and the rest as the unlabeled data, respectively. In the supplement, we provide more results for varying numbers of unlabeled data. To demonstrate how VL-PLM improves PLs for SSOD, we mainly compare our method with the following baselines. (1) *Supervised*: A vanilla teacher model trained on the labeled set $\mathcal{I}_L$. (2) *Supervised*+PLs: We apply the vanilla teacher model on the unlabeled set $\mathcal{I}_U$ to generate PLs and train a student model with both ground truth and PLs. To compare with *Supervised*+PLs, VL-PLM generates PLs for all categories on $\mathcal{I}_U$. Then, those PLs are merged into the PLs from the vanilla teacher as the final PLs to train a student model named as *Supervised*+VL-PLM. (3) STAC [44]:

Table 3. Evaluation of pseudo labels for semi-supervised object detection on COCO [32].

Methods	1% COCO	2% COCO	5% COCO	10% COCO
Supervised	9.25	12.70	17.71	22.10
Supervised+PLs	11.18	14.88	21.20	25.98
Supervised+VL-PLM	**15.35**	**18.60**	**23.70**	**27.23**
STAC [44]	13.97	18.25	24.38	28.64
STAC+VL-PLM	**17.71**	**21.20**	**26.21**	**29.61**

A popular SSOD baseline. To compare with STAC, we only replace its PLs with ours that are used to train *Supervised*+VL-PLM. The new STAC student model is denoted as STAC+VL-PLM. Here we report the standard metric for COCO, mAP, which is an average over IoU thresholds from 0.5 to 0.95 with a step size of 0.05.

Implementation details: We follow the same PL generation pipeline and hyperparameters as the OVD experiment, except that we take a class-agnostic Faster R-CNN [40] as our proposal generator and train it on the different COCO splits. *Supervised* and *Supervised*+PLs are implemented in Detectron2 [48] and trained for 90,000 iterations with a batch size of 16. For models related to STAC [44], we use the official code of STAC with default settings.

Results: As shown in Table 3, models with VL-PLM outperform *Supervised* + PLs and STAC by a clear margin, respectively. Since the only change to the baselines is the addition of VL-PLM's PLs, we can conclude that V&L adds clear value to the PLs and can benefit SSOD. Another interesting finding is that models with VL-PLM provide bigger gains for smaller labeled data, which is the most important regime for SSOD as it brings down annotation costs. In that regime, PLs from V&L models are likely stronger than PLs from the small amount of annotated data. We also want to mention two recent SSOD methods [49,60] that achieve higher absolute performance, however, only with additional and orthogonal contributions. VL-PLM may also improve these methods, but here we focus on a fair comparison to other PL-based methods. Moreover, we believe that with better V&L models, VL-PLM can further improve SSOD.

4.3 Analysis of pseudo label generation

We base our ablation studies on the COCO-ZS setting for OVD unless otherwise specified. All models are trained for 90,000 iterations with a batch size of 16.

Understanding the quality of PLs: Average precision (AP) is a dominant metric to evaluate object detection methods. However, AP alone does not fully indicate the quality of PLs, and the number of PLs also needs to be considered. To support this claim, we generate 5 sets of PLs as follows. (1) *PL v1*: We take the raw region proposals from RPN without RoI refinement in our pseudo label generation and set $\tau = 0.05$. (2) *PL v2*: The same as *PL v1* but with $\tau = 0.95$. (3) *PL v3*: VL-PLM with $\tau = 0.05$. (4) *PL v4*: VL-PLM with $\tau = 0.95$. (5)

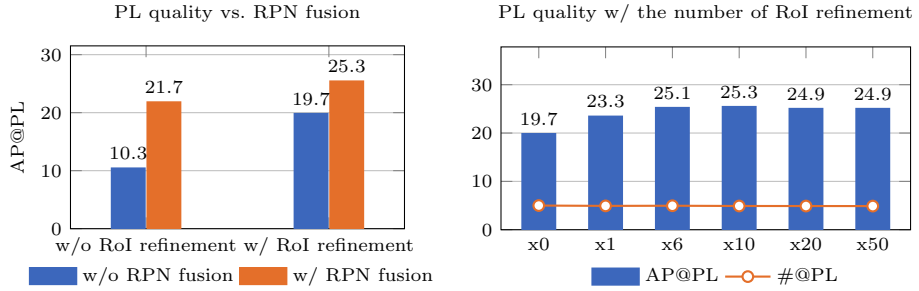

Fig. 4. The quality of PLs with different combinations of RPN and RoI head. We change the threshold τ to ensure each combination with a similar #@PL. "×N" means we apply RoI head N times to refine the proposal boxes.

Table 4. Relationship between the quality of pseudo labels and the performance of the final open vocabulary detectors.

	PL Setting	Pseudo Labels		Final Detector		
		AP@PL	#@PL	Base AP	Novel AP	Overall AP
PL v1	No RoI, $\tau = 0.05$	17.4	89.92	33.3	14.6	28.4
PL v2	No RoI, $\tau = 0.95$	14.6	2.88	56.1	26.0	48.2
PL v3	VL-PLM, $\tau = 0.05$	20.6	85.15	29.7	19.3	27.0
PL v4	VL-PLM, $\tau = 0.95$	18.0	2.93	55.4	**31.3**	**49.1**
PL v5	VL-PLM, $\tau = 0.99$	11.1	1.62	**56.7**	27.2	49.0

PL v5: VL-PLM with $\tau = 0.99$. In Table 4, we report AP_{50} (AP@PL) and the average per-image number (#@PL) of pseudo labels on novel categories. We also report the performance of detection models trained with the corresponding PLs as Novel AP, Base AP and Overall AP. Comparing *PL v1* with *PL v4* and *PL v2* with *PL v4*, we can see that a good balance between AP@PL and #@PL is desired. Many PLs may achieve high AP@PL, but drop the performance of the final detector. A high threshold reduces the number of PLs but degrades AP@PL as well as the final performance. We found $\tau = 0.8$ to provide a good trade-off. The table also demonstrates the benefit of VL-PLM over no RoI refinement. The supplement contains more analysis and visualizations of our pseudo labels.

Two-stage proposal generator matters: As mentioned in Sec. 3.2, we improve the localization ability of CLIP with the two-stage proposal generator in two ways: 1) we merge CLIP scores with RPN scores, and 2) we repeatedly refine the region proposals from RPN with the RoI Head. To showcase how RPN and the RoI head help PLs, we evaluate the quality of PLs from different settings in Fig. 4. As shown, RPN score fusion always improves the quality of PLs. As we increase the number of refinement steps with RoI head, the quality increases and converges after about 10 steps. Besides proposals from our RPN with RoI

Table 5. The quality of pseudo labels generated from different region proposals. The threshold τ is tuned to ensure a similar #@PL for each method.

	Selective Search [46]	RoI Head	RPN	RPN+RoI (Ours)
τ	0.99	0.55	0.88	0.82
AP@PL	5.7	8.8	19.7	**25.3**
#@PL	34.92	5.01	4.70	**4.26**

refinement (RPN+RoI), we investigate region proposals from different sources, i.e. 1) Selective search [46], 2) RPN only, and 3) RoI head with default thresholding and NMS. Table 5 shows that selective search with a high τ still leads to a large #@PL with a low AP@PL for at least two reasons. First, unlike RPN, selective search does not provide objectiveness scores to improve the localization of CLIP. Second, it returns ten times more proposals than RPN, which contain too many noisy boxes. Finally, the RoI head alone also leads to a poor quality of PLs because it classifies many novel objects as background, due to its training protocol. In the supplement, we show that the proposal generator, which is trained on base categories, generalizes to novel categories.

Time efficiency: VL-PLM sequentially generates PLs for each region proposal, which is time-consuming. For example, VL-PLM with ResNet50 takes 0.54s per image on average. We provide two solutions to reduce the time cost. 1) Simple multithreading on 8 GPUs can generate PLs for the whole COCO training set within 6 hours. 2) We provide a faster version (Fast VL-PLM) by sharing the ResNet50 feature extraction for all region proposals of the same image. This reduces the runtime by 5× with a slight performance drop. Adding multi-scale features (Multiscale Fast VL-PLM) avoids the performance drop but still reduces runtime by 3×. Please refer to the supplement for more details.

5 Conclusion

This paper demonstrates how to leverage pre-trained V&L models to mine unlabeled data for different object detection tasks, e.g., OVD and SSOD. We propose a V&L model guided pseudo label mining framework (VL-PLM) that is simple but effective, and is able to generate pseudo labels (PLs) for a task-specific labelspace. Our experiments showcase that training a standard detector with our PLs sets a new state-of-the-art for OVD on COCO. Moreover, our PLs can benefit SSOD models, especially when the amount of ground truth labels is limited. We believe that VL-PLM can be further improved with better V&L models.

Acknowledgments. This research has been partially funded by research grants to D. Metaxas from NEC Labs America through NSF IUCRC CARTA-1747778, NSF: 1951890, 2003874, 1703883, 1763523 and ARO MURI SCAN.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV (2019)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)
4. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: ECCV. pp. 384–400 (2018)
5. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR (2018)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: ECCV (2020)
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server (2015)
8. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. In: ECCV (2020)
9. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., , Batra, D.: Embodied Question Answering. In: CVPR (2018)
10. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: ICCV. pp. 2876–2885 (2021)
11. Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
12. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From Captions to Visual Concepts and Back. In: CVPR (2015)
13. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In: EMNLP (2016)
14. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Towards open vocabulary object detection without human-provided bounding boxes (2021)
15. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR. pp. 2918–2928 (2021)
16. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In: ICLR (2022)
17. Gupta, A., Dollár, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
19. Hu, R., Singh, A.: UniT: Multimodal Multitask Learning with a Unified Transformer. In: ICCV (2021)
20. Hudson, D.A., Manning, C.D.: Learning by Abstraction: The Neural State Machine. In: NeurIPS (2019)
21. Huynh, D., Kuen, J., Lin, Z., Gu, J., Elhamifar, E.: Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling (2021)

22. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In: CVPR (2018)
23. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: ICML (2021)
24. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. In: ICCV (2021)
25. Karpathy, A., Fei-Fei, L.: Deep Visual-Semantic Alignments for Generating Image Descriptions. In: CVPR (2015)
26. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to Objects in Photographs of Natural Scenes. In: EMNLP (2014)
27. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)
28. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven Semantic Segmentation. In: ICLR (2022)
29. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)
30. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In: ECCV (2020)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: CVPR (2017)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
33. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
34. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: NeurIPS (2019)
35. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: CVPR (2016)
36. Peng, G., Jiang, Z., You, H., Lu, P., Hoi, S., Wang, X., Li, H.: Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. In: CVPR (2019)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
38. Rahman, S., Khan, S., Barnes, N.: Improved visual-semantic alignment for zero-shot object detection. In: AAAI. pp. 11932–11939 (2020)
39. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting (2021)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: NeurIPS (2015)
41. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Li, J., Zhang, X., Sun, J.: Objects365: A Large-scale, High-quality Dataset for Object Detection. In: ICCV (2019)
42. Shi, H., Hayat, M., Wu, Y., Cai, J.: Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues (2022)

43. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. In: BMVC (2021)
44. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. In: arXiv:2005.04757 (2020)
45. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
46. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV (2013)
47. Wang, L., Li, Y., Lazebnik, S.: Learning Deep Structure-Preserving Image-Text Embeddings. In: CVPR (2016)
48. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
49. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: ICCV. pp. 3060–3069 (2021)
50. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model (2021)
51. Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., Chen, X.: Unsupervised Domain Adaptation for Object Detection via Cross-Domain Semi-Supervised Learning. In: WACV (2022)
52. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., L.Berg, T.: MAttNet: Modular Attention Network for Referring Expression Comprehension. In: CVPR (2018)
53. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling Context in Referring Expressions. In: ECCV (2016)
54. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-Vocabulary Object Detection Using Captions. In: CVPR (2021)
55. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting Visual Representations in Vision-Language Models. In: CVPR (2021)
56. Zhao, X., Schulter, S., Sharma, G., Tsai, Y.H., Chandraker, M., Wu, Y.: Object Detection with a Unified Label Space from Multiple Datasets . In: ECCV (2020)
57. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining (2021)
58. Zhong, Y., Wang, J., Peng, J., Zhang, L.: Boosting weakly supervised object detection with progressive knowledge transfer. In: ECCV. pp. 615–631. Springer (2020)
59. Zhou, C., Loy, C.C., Dai, B.: Denseclip: Extract free dense labels from clip (2021)
60. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In: CVPR (2021)
61. Zhu, P., Wang, H., Saligrama, V.: Don't even look once: Synthesizing features for zero-shot detection. In: CVPR. pp. 11693–11702 (2020)