LP-OVOD: Open-Vocabulary Object Detection by Linear Probing

Chau Pham* Truong Vu* Khoi Nguyen
VinAI Research

Abstract

*This paper addresses the challenging problem of open-vocabulary object detection (OVOD) where an object detector must identify both seen and unseen classes in test images without labeled examples of the unseen classes in training. A typical approach for OVOD is to use joint text-image embeddings of CLIP to assign box proposals to their closest text label. However, this method has a critical issue: many low-quality boxes, such as over- and under-covered-object boxes, have the same similarity score as high-quality boxes since CLIP is not trained on exact object location information. To address this issue, we propose a novel method, LP-OVOD, that discards low-quality boxes by training a sigmoid linear classifier on pseudo labels retrieved from the top relevant region proposals to the novel text. Notably, LP-OVOD seamlessly integrates the knowledge distillation technique from ViLD, resulting in a new state-of-the-art OVOD approach. Experimental results on COCO affirm the superior performance of our approach over prior work, achieving **40.5** in AP$_{novel}$ using **ResNet50** as the backbone and without external datasets or knowing novel classes in training. Our code will be available at `https://github.com/VinAIResearch/LP-OVOD`.*

1. Introduction

Open-vocabulary object Detection (OVOD) is an important and emerging computer vision problem. The task is to detect both seen and unseen classes in test images, given only bounding box annotations of seen classes in the training set. Seen classes are called base classes while unseen classes are called novel classes, both explicitly specified by their names. Novel classes are determined based on the availability of annotations for those classes in the training set. Classes present in training images without annotations are still considered novel classes. OVOD has various applications where a detector should be capable of extending its detected categories to novel classes without human annotation such as in autonomous driving or augmented reality where new classes can appear in deployment without annotation. OVOD is also useful as an automatic labeling system in scenarios where it is impractical for annotators to exhaustively label all objects of all classes in a large dataset.

Figure 1. Comparison of box predictions for novel classes 'bus' and 'cake' between ViLD [10] (top) and our approach (bottom). Low-quality boxes have similar scores to high-quality ones, leading to high false positive (left) and false negative rates (right) in ViLD while being eliminated in our LP-OVOD.

The main challenge in OVOD is to detect novel classes without labels while maintaining good performance for base classes. To address this challenge, a pretrained visual-text embedding model, such as CLIP [28] or ALIGN [15], is provided as a joint text-image embedding where base and novel classes co-exist. This embedding can be used to align box proposals with their closest classes. However, for each object in the image, low-quality (over- and under-covered-object boxes) and high-quality (perfectly match the objects' extent) box proposals can co-exist as they have the same similarity scores to the text embeddings. This is because CLIP is trained on images without object location information, leading to high false positive and false negative rates in the OVOD approaches as shown in Fig. 1.

To address this limitation, we propose a novel linear probing method called LP-OVOD. This technique involves training a linear classifier for novel classes using features from the penultimate layer of a Faster R-CNN model pretrained on base classes. Despite being trained solely on

*The first two authors contribute equally.

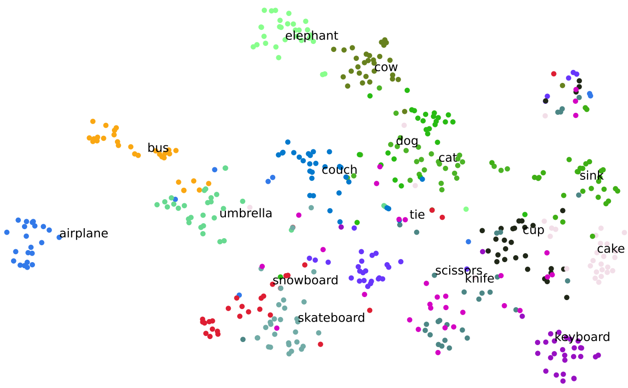

Figure 2. The feature embeddings of COCO novel classes are extracted from the penultimate layer of a Faster R-CNN pretrained on base classes. These embeddings are highly discriminative, which motivates us to learn a robust classifier from them.

base classes, these features are remarkably discriminative for novel classes (Fig. 2). To generate pseudo labels for the linear classifier, we extract box candidates from the most relevant proposal boxes associated with the novel text. This way, our approach effectively harnesses related classes within training images, even without annotations. These related classes include exact and similar classes like 'kitten' and 'cat', or 'pony' and 'horse'. Importantly, we do not assume novel class presence in training images.

Additionally, for smooth integration with the linear classifier covering both base and novel classes, without manual score adjustment, we propose using a sigmoid classifier. The sigmoid classifier predicts each class independently, making it conducive to swift combination. By concatenating the weights of the novel class linear classifier with those of the base class linear classifier, our approach facilitates object classification across both base and novel classes. This classification score, combined with the distillation score from ViLD's knowledge distillation head [10], significantly enhances the model's ability to differentiate coarse-grained and fine-grained classes.

We demonstrate the effectiveness of our approach on two standard OVOD datasets: COCO [22] and LVIS [11]. LP-OVOD significant improvement over state-of-the-art methods, without relying on external datasets or retraining the whole network whenever novel classes arrive.

In summary, the contributions of our work are as follows:

- A linear probing approach that leverages the highly discriminative features extracted from the penultimate layer of a pretrained Faster R-CNN on base classes to train a linear classifier for novel classes on the pseudo labels from retrieving the top relevant box proposals.
- Sigmoid classifiers for both pretraining on base classes and linear probing on novel classes to predict class scores independently, forming a unified classifier for both base and novel classes in testing.

In the following, Sec. 2 reviews prior work; Sec. 3 specifies our approach; and Sec. 4 presents our experimental results. Sec. 5 concludes with some remarks.

2. Related Work

Object detection approaches aiming to localize and classify objects in images can be classified into three groups: anchor-based, anchor-free, and DETR-based detectors. Anchor-based detectors, such as Faster RCNN [32], RetinaNet [21], and YOLO [31], first classify and then regress the predefined anchor boxes. In contrast, anchor-free detectors like CenterNet [48] and FCOS [34] regress the bounding box extent directly without using predefined anchor boxes. DETR-based detectors [3, 19, 23, 37, 44, 50] leverage encoder-decoder transformer architecture along with one-to-one matching loss to predict object bounding boxes in an end-to-end manner without using NMS. However, these methods are designed to work in a closed-vocabulary setting, where detectors are trained and evaluated on predefined categories, unlike our OVOD setting.

Few-shot object detection (FSOD) approaches [7, 27, 36, 41] aim to detect novel objects with a few labeled examples. On the other hand, OVOD only requires the names of the novel classes instead. These two inputs are complementary since some fine-grained classes may be easier to identify through exemplars, while others may be more common and easier to identify through their names.

Zero-shot or open-vocabulary object detection (ZSOD/OVOD) aims to detect unseen categories given the class name. To enable open-vocabulary learning, during training, we are provided with labeled examples of the base classes and a pretrained word embedding (such as Word2vec [24], GloVe [26]), or vision-language models (such as CLIP [28], ALIGN [15]). OVOD methods can be grouped as follows:

External-dataset-based methods [2, 8, 9, 14, 20, 25, 30, 35, 39, 43, 46, 47] utilize huge external datasets, including image-caption pairs or image-level labeled annotations, to improve the pretrained vision-language model or detectors to recognize more classes, including the novel ones. Thus, these methods have an advantage over those that do not.

Novel-class-aware methods including OV-DETR [42], VL-PLM [45] assume that novel categories are known during training. These methods retrieve large-scale region proposals of novel classes based on the joint text-image embedding of CLIP [28] as pseudo-GT labels, which are jointly trained with GT-labeled examples of base classes. As a result, these methods need to regenerate the pseudo labels and retrain the detectors whenever new classes arrive.

Novel-class-unaware methods [5, 10, 18, 40] follow the same setting as ours. ViLD [10] uses knowledge distillation from CLIP visual features to learn the embedding for unseen categories. DetPro [5] proposes a learnable-text

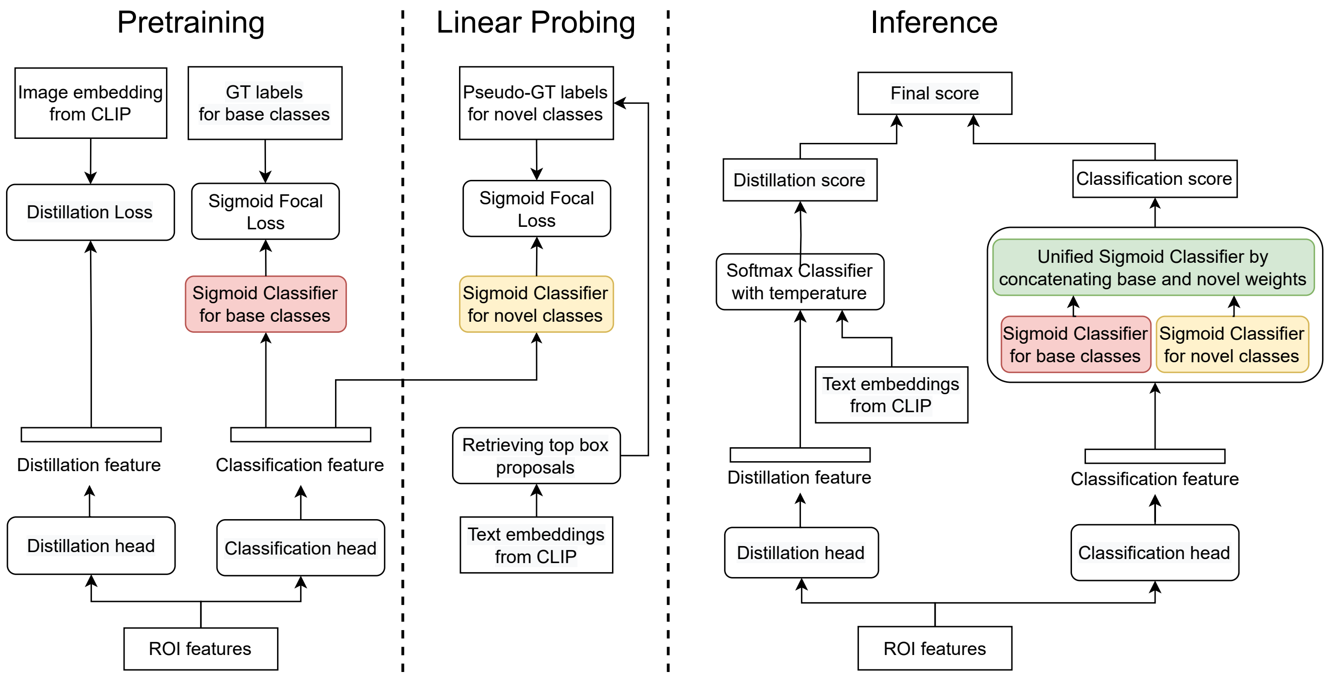

Figure 3. **Overview of our approach.** LP-OVOD starts with extracted ROI features from Faster R-CNN [32], following the initial steps. During pretraining **(left)**, a distillation head mimics CLIP's predictions, like ViLD [10]. We replace the softmax classifier with a sigmoid variant, trained using base class labels. In the linear probing phase **(middle)**, a novel sigmoid classifier with a trainable linear layer is trained using pseudo labels from novel classes. These labels are derived from top box proposals based on novel text embeddings. In inference **(right)**, we concatenate both sigmoid classifier weights for a unified classifier spanning base and novel classes. This unified classifier independently predicts class scores. Final detection scores result from combining classification and distillation scores.

prompt instead of a fixed-text prompt. F-VLM [18] utilizes a pretrained CLIP's image encoder as a backbone to retain the locality-sensitive features necessary for detection.

However, these methods attempt to align the text embedding with the feature embedding of each proposal to predict its class. In contrast, our method approaches a different way that learns a linear classifier for novel classes using features extracted from a Faster R-CNN pretrained on base classes.

3. Our Approach

Problem statement: During training, we are provided with a large set of annotated examples of base classes C_B, i.e., bounding boxes b_i and their categories $c_i \in C_B$. In testing, given the names of novel classes C_N, our goal is to detect objects of both base and novel classes, i.e., $\hat{c}_i$, $\hat{b}_i$, where $\hat{c}_i \in C_B \cup C_N$ for test images. To facilitate learning, a pretrained CLIP [28] is provided as the joint image-text embedding of both base and novel classes.

Our scope: Our approach strictly assumes that we do not know novel classes during training, as we cannot anticipate the classes that an open-vocabulary detector (OVD) will encounter in practical use. Additionally, to ensure a fair comparison, we utilize only the images and annotations provided by each benchmark without any external datasets, such as image-caption or image-level label datasets.

Fig. 3 illustrates our approach, which is based on Faster R-CNN [32]. We adopt the same backbone, region proposal network (RPN), and box regression modules, and refer readers to [32] for details. However, we make two modifications: replacing the softmax classifier with a sigmoid classifier and adding a distillation head as in ViLD [10]. For novel classes, we extract features from the top relevant proposals to the novel text embedding as pseudo labels for training a sigmoid classifier of the novel classes. In testing, we concatenate the weights of the two sigmoid classifiers to form a unified sigmoid classifier for object detection.

3.1. Pretraining on Base Classes

To accelerate learning of novel classes during testing, we propose substituting Faster R-CNN's softmax classifier [32] with a sigmoid variant. This pretrains the classifier on base classes, transforming the task from multi-category classification to detecting the presence or absence of a category within an image. This prevents new class embeddings from clustering with base class embeddings, as shown in Fig. 2. This classifier also independently predicts each category. This design enables seamless integration of newly trained classifier weights with those of base classes, creating a unified classifier for both base and new classes. This eliminates

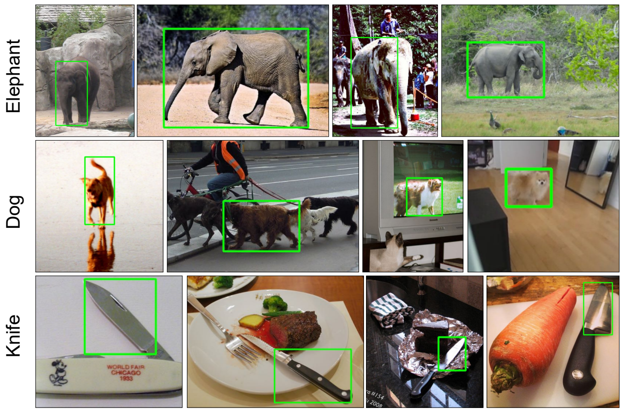

Figure 4. Top-4 box proposal retrievals from CLIP's embeddings of four novel classes: 'elephant', 'dog', and 'knife'. The quality is good enough to be used as pseudo labels for training a few-shot classifier on novel classes.

the need for retraining or temperature adjustments.

Concretely, for each proposal $\tilde{b}_i$, we extract ROI features from the backbone. These features are used by the classification and distillation heads to obtain classification feature f_i^{cls} and distillation feature f_i^{dis}, respectively. We jointly train a new sigmoid classifier and the distillation head. The sigmoid classifier learns from base class ground-truth labels c_i using sigmoid focal loss [21]. Simultaneously, the distillation head is guided by CLIP's image embedding e_i^{image}, which comes from cropped images of proposal $\tilde{b}_i$. The distillation head is trained using L1 loss. In particular,

$$\mathcal{L}_{\text{cls}}^{\text{Base}} = \sum_i \textbf{Focal loss}(\text{Sigmoid}(f_i^{\text{cls}}; W_B), c_i), \quad (1)$$

$$\mathcal{L}_{\text{dis}} = \sum_i \| f_i^{\text{dis}} - e_i^{\text{image}} \|_1, \quad (2)$$

where W_B are the weights of the base classes.

3.2. Linear Probing on Novel Classes

As illustrated in Fig. 1, low-quality boxes usually have the same similarity score to the novel text embeddings as the high-quality ones do, resulting in high false positive and false negative rates. Therefore, we need to have better positive/negative proposals for training a sigmoid classifier to discard these low-quality proposals.

To this end, first, the top relevant proposals of each novel class are retrieved and served as pseudo-GT labels $\tilde{c}_i$. Specifically, we extract all image embeddings e_i^{image} of all proposals $\tilde{b}_i$ having the objectness score o_i larger than τ in the training set. For each novel category with text embedding e_c^{text} where $c \in C_N$, we retrieve the top K closest proposals in order to form a set $\mathcal{P} = \{(\tilde{b}_i, \tilde{c}_i)\}_{i=1..K \times C_N}$ using cosine similarity $\cos(e_c^{\text{text}}, e_i^{\text{image}})$. We visualize the examples of top-4 retrieved proposal for four novel classes in Fig. 4. To speed up the retrieval process, we resort the Faiss [16] tool. Then, we leverage the sampling mechanism of Faster R-RCNN to sample positive/negative proposals where the positives $\mathcal{P}^+ = \{(\tilde{b}_i, \tilde{c}_i)\}, \tilde{c}_i \in N_c$ are the ones having IoU > 0.5 with the pseudo-GT boxes $\mathcal{P}$ while the rest are the negatives $\mathcal{P}^- = \{(\tilde{b}_i, 0)\}$.

When novel classes arrive, a new sigmoid classifier W_N is added on top of the pretrained classification feature f_i^{cls}. The sigmoid classifier for novel classes is trained as follows:

$$\mathcal{L}_{\text{cls}}^{\text{Novel}} = \sum_{i=1}^{|\mathcal{P}^+ \cup \mathcal{P}^-|} \textbf{Focal loss}(\text{Sigmoid}(f_i^{\text{cls}}; W_N), c_i), \quad (3)$$

where W_N are the weights of the novel classes.

Discussion: Our approach achieves speed, requiring only 5 minutes on COCO, as we focus solely on retrieving top proposals. This differs from OV-DETR [42] and VL-PLM [45], which extract pseudo labels from the entire training set for joint training with base class labels. Moreover, our approach does not hinge on acquiring precise proposals for the novel text. This adaptability arises because our model proficiently classifies fine-grained classes, aided by the distillation head. Consequently, proposals from related classes also yield strong results, minimizing the need for precise novel class proposals, even in scenarios where training data might lack examples of novel classes.

3.3. Inference on Both Base and Novel Classes

Given a proposal box $\tilde{b}_i$ with classification feature f_i^{cls} and distillation feature f_i^{dis}, the inference on both base and novel classes is visualized in the right of Fig. 3.

For the classification head, we concatenate the weights of the sigmoid classifiers learned on the base and novel classes to form a unified classifier with weight $W = [W_B; W_N]$. The classification score s_i^{cls} is calculated as:

$$s_i^{\text{cls}} = \text{Sigmoid}(f_i^{\text{cls}}; W) \in [0,1]^{|C_B|+|C_N|}. \quad (4)$$

For the distillation head, we compute the distillation score s_i^{dis} as the softmax score of the cosine similarity between the distillation features f_i^{dis} and text embeddings e_c^{text} with temperature κ as:

$$s_i^{\text{dis}} = \text{Softmax}_c \left(\frac{\cos(f_i^{\text{dis}}, e_c^{\text{text}})}{\kappa} \right) \in [0,1]^{|C_B|+|C_N|}. \quad (5)$$

Finally, the final score for prediction of each proposal $\tilde{b}_i$ with objectness score o_i is computed as:

$$s_i = o_i \cdot \begin{cases} s_i^{\text{cls}} \text{ for base classes} \\ (s_i^{\text{cls}})^{\beta} (s_i^{\text{dis}})^{1-\beta} \text{ for novel classes} \end{cases} \quad (6)$$

where β are coefficient hyper-parameter for novel classes.

Method	Venue	Training source	Box AP on COCO			Mask AP on LVIS			
			AP$_{novel}$	**AP$_{base}$**	**AP**	**AP$_r$**	**AP$_f$**	**AP$_c$**	**AP**
OVR-CNN [43]	CVPR 21	image captions in $C_B \cup C_N$ instance-level labels in C_B (use external datasets)	22.8	46.0	39.9	-	-	-	-
XPM [14]	CVPR 22		27.0	46.3	41.3	-	-	-	-
RegionCLIP [46]	CVPR 22		31.4	57.1	50.4	17.1	27.4	34.0	28.2
PromptDet [8]	ECCV 22		26.6	59.1	50.6	19.0	18.5	25.8	21.4
Detic [47]	ECCV 22		27.8	47.1	42.1	17.8	26.3	31.6	26.8
PB-OVD [9]	ECCV 22		30.8	46.1	30.1	-	-	-	-
OWL-ViT [25]	ECCV 22		41.8	49.1	47.2	16.9	-	-	19.3
VLDet [20]	ICLR 23		32.0	50.6	45.8	21.7	29.8	34.3	30.1
MS-OVIS [35]	CVPR 23		31.5	-	-	-	-	-	-
BARON [39]	CVPR 23		42.7	54.9	51.7	22.6	27.6	29.8	27.6
OV-DETR [42]	ECCV 22	instance-level labels in C_B known novel classes during training	29.4	61.0	52.7	17.4	25.0	32.5	26.6
VL-PLM [45]	ECCV 22		34.4	60.2	53.5	17.2	23.7	35.1	27.0
ZSD [1]	ECCV 18	instance-level labels in C_B (zero-shot object detection)	0.31	29.2	24.9	-	-	-	-
PL [29]	AAAI 20		4.12	35.9	27.9	-	-	-	-
DELO [49]	CVPR 20		3.41	13.8	11.1	-	-	-	-
ViLD [10]	ICLR 22	instance-level labels in C_B unknown novel classes during training	27.6	59.5	51.2	16.6	24.6	30.3	25.5
RegionCLIP† [46]	CVPR 22		14.2	52.8	42.7	-	-	-	-
DetPro‡ [5]	CVPR 22		19.8	60.2	49.6	<u>19.8</u>	25.6	28.9	25.9
F-VLM [18]	ICLR 23		28.0	43.7	39.6	18.6	-	-	24.2
CORA [40]	CVPR 23		35.1	35.5	35.4	-	-	-	-
LP-OVOD (ours)	-	(with OLN [17] proposals)	<u>40.5</u>	60.5	<u>55.2</u>	19.3	26.1	29.4	<u>26.2</u>
LP-OVOD (ours)	-	(with OWL-VIT [25] proposals)	**44.9**	59.4	**55.6**	**23.0**	28.0	30.4	**28.6**

Table 1. **Performance on COCO and LVIS.** '-' indicates unreported number. † is an alternate RegionCLIP version using only COCO for training. ‡ is our COCO DetPro re-run, excluding LVIS transfer. Methods in faded rows are for reference only. **Best** results, <u>second best</u> .

4. Experimental Results

Datasets: We conduct our experiments using the OVOD versions called OV-COCO [1] and OV-LVIS [10] of two public datasets: COCO [22] and LVIS [11]. The OV-COCO dataset comprises 118,000 images with 48 base categories and 17 novel categories. OV-LVIS [11] shares the image set with OV-COCO. Its categories are divided into 'frequent', 'common', and 'rare' groups based on their occurrences, representing the long-tailed distributions of 1,203 categories. We treated the 'frequent' and 'common' groups of 866 categories as the base classes while considering the rare' group of 337 categories as the novel classes.

Evaluation metrics: Consistent with the standard OVOD evaluation protocol [10,46], we report the box Average Precision (AP) with an IoU threshold of 0.5 for object detection on the COCO dataset, i.e. AP$_{novel}$ for novel classes, AP$_{base}$ for base classes, and AP for all classes. For instance segmentation on the LVIS dataset, we report the mask AP, which is the average AP over IoU thresholds ranging from 0.5 to 0.95, i.e., AP$_r$, AP$_f$, AP$_c$, and AP for 'rare', 'frequent', 'common', and all classes, respectively.

Implementation details: In our implementation, we use the Faster R-CNN detector [32] for COCO and the Mask-RCNN detector [12] for LVIS, both with the ResNet50 [13] backbone. The ResNet50 backbone is initialized with the self-supervised pre-trained SoCo [38]. We use multi-scale training with different image sizes while maintaining the aspect ratio for data augmentation. We employ OLN [17] as the object proposal network. For training on base classes, we use the SGD optimizer with an initial learning rate of 0.02 and an image batch size of 16. We adopt the 20-epoch schedule from MMDetection [4], where the learning rate is decreased by a factor of 10 after the 16th and 19th epochs, and apply a linear warm-up learning rate for the first 500 iterations. For quick adapting to novel classes, we set the objectness score threshold to $\tau = 0.6$ to filter proposals before retrieval. We train the novel weights W_N for 12 epochs using the SGD optimizer with an initial learning rate of 0.01 and decreasing the learning rate by a factor of 10 after the 8th and 11th epochs. In testing, we use a temperature of $\kappa = 0.01$ for the distillation head.

4.1. Comparison with State-of-the-art Approaches

Results on COCO are shown in Tab. 1 and Fig. 5. In Tab.1, we compare our approach to diverse methods: ZSOD, external-data-based, novel-class-aware, and novel-class-unaware. Our approach significantly outperforms the second-best (CORA) on COCO by +5.5 in AP$_{novel}$, while maintaining a much better base class performance (60.5 vs.

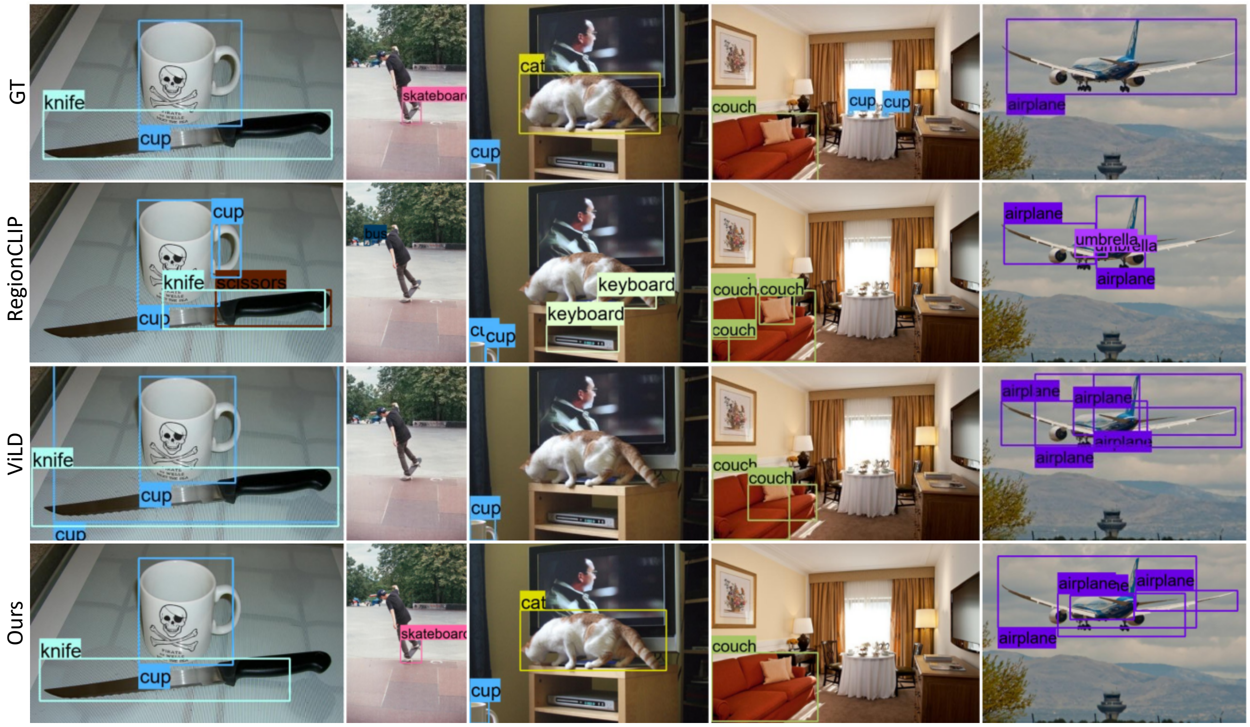

Figure 5. **Qualitative comparison of different approaches on COCO's novel classes.** The first four columns show our superior performance while the last one shows a failure case where all of them cannot generate boxes for the airplane due to its rare aspect ratio.

	AP$_{novel}$ on COCO	**AP$_r$ on LVIS**
Ours + RPN [32]	37.2	19.3
Ours + OLN [17]	**40.5**	19.3

Table 2. The effectiveness of RPN [32] and OLN [17].

Retrieval	**Sigmoid**	**AP$_{novel}$**	**AP$_{base}$**	**AP**
		27.6	61.2	52.4
✓		33.2	**61.2**	53.9
✓	✓	**40.5**	60.5	**55.2**

Table 3. Ablation study on the contribution of each component. **Retrieval**: retrieving top boxes as pseudo labels for novel classes. **Sigmoid**: replace softmax with sigmoid classifier.

35.5). Notably, LP-OVOD far exceeds ViLD [10] (our baseline) by 13 in AP$_{novel}$, with only a new classification head. In Fig. 5, our approach excels. RegionCLIP [46] misclassifies foreground instances, and ViLD [10] generates redundancies. The last column showcases a challenge where methods struggle due to airplane aspect ratio. These results affirm our effectiveness without external data or known novel classes in training.

Results on LVIS are shown in Tab. 1 and Fig. 6 reveals that our results are comparable to DetPro [5], yet the improvement is less pronounced than in COCO. This discrepancy stems from the semantic distinction between base and novel classes. In COCO, the difference is relatively high due to fewer classes, resulting in coarser granularity. In contrast, LVIS classes are finer-grained, facilitating easier transfer of learned embeddings from base to novel classes. This allows swift matching of novel text embeddings with predicted features during testing, diminishing the significance of our linear probing method in this context. Nonetheless, we still significantly outperform our baseline, ViLD [10], by approximately 3 in AP$_r$, courtesy of our proposed classification head.

4.2. Ablation Study

In this section, we conduct ablation studies on the COCO dataset on various aspects to analyze our approach.

Impact of different proposal networks. Tab. 2 presents the results of our approach using RPN [32] and OLN [17] proposals. OLN is a SOTA object proposal network in the open-world setting. On COCO, the quality of the OLN proposals is higher than that of RPN with the same supervision in training, as evidenced by an improvement of +3.3 in AP$_{novel}$. This is because OLN is more robust to object sizes and aspect ratios by replacing foreground/background classification with centerness and IoU score predictions. However, when the number of base classes increases, as in the

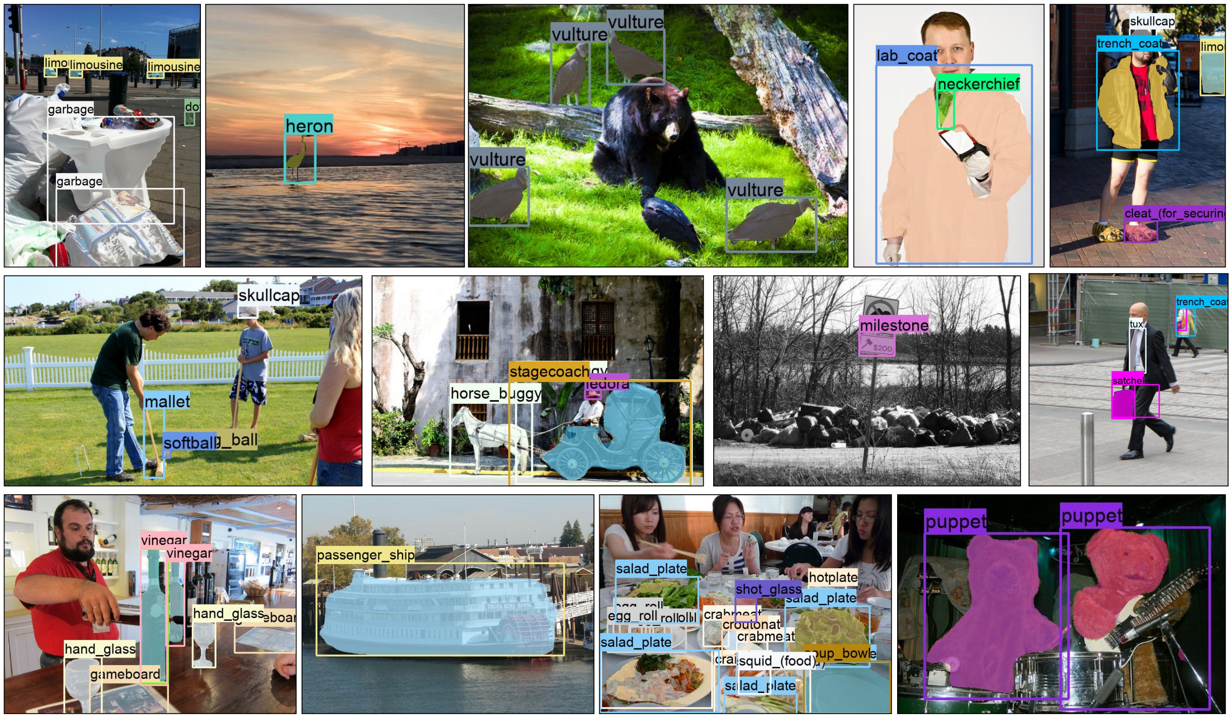

Figure 6. **Qualitative results of novel classes on the LVIS dataset [11].** Our approach can successfully detect some novel classes including "lab coat", "mallet", and "hand glass". However, due to the rarity of some novel classes in training, our method retrieves the proposals of close-meaning classes instead, i.e., "tie" vs "neckerchief", leading to the wrong prediction in testing.

Features	AP$_{novel}$	AP$_{base}$	AP
Classification	**35.9**	60.5	**54.1**
Distillation	19.7	60.5	49.8

Table 4. Types of features to learn the sigmoid linear classifier.

# proposals	5	10	20	50	100	200
AP$_{novel}$	30.5	34.8	38.3	40.3	**40.5**	39.6

Table 5. Ablation on # retrieved proposals per novel class.

case of LVIS, these predictions become less effective since the base classes can cover a wider range of object sizes and aspect ratios of the novel classes.

Ablation study on each component's contribution is summarized in Tab. 3. Our baseline is ViLD with OLN proposals. By using retrieval of top boxes as the pseudo labels for novel classes, the performance improves significantly by +5.6 in AP$_{novel}$ compared to the baseline, while keeping the performance of base classes intact. Moreover, combining the sigmoid classifier and the pseudo-labeling strategy results in the best performance of 40.5 in AP$_{novel}$.

Study on features to learn the sigmoid classifier. To quantitatively show that the classification features of Faster R-CNN pre-trained on base classes are superior to distillation

β	0.9	0.8	0.7	0.6
AP$_{novel}$	40.2	**40.5**	39.7	38.5

Table 6. Study on the coefficient of novel classes β.

features for classifying novel classes, we train a sigmoid classifier on top of the classification feature f_i^{cls} and the distillation feature f_i^{dis}, which is trained to distill the CLIP's image embedding. The results are presented in Tab. 4. The feature of the classification head yields 35.9 in AP$_{novel}$, greatly outperforming that of the distillation head.

Number of retrieved proposals per novel class. Tab. 5 presents the performance of our approach for different numbers of proposals K per novel class in Sec. 3.2. The performance improves as the value of K increases and saturates at K=100. We speculate that a higher number of proposals provides more diverse examples for training whereas too many proposals increase the likelihood of including noisy boxes, resulting in suboptimal performance. Moreover, too many proposals can slow down the retrieval and few-step training of the linear classifier for novel classes.

Study on the coefficient of novel classes β is summarized in Tab. 6. ViLD uses $\beta = 1/3$, indicating that the distillation head's novel scores have more impact on the final prediction than the classification head's scores. However,

Objectness	AP$_{novel}$	AP$_{base}$	AP
	34.6	**61.4**	54.4
✓	**40.5**	60.5	**55.2**

Table 7. The importance of objectness score o_i in Eq. (6).

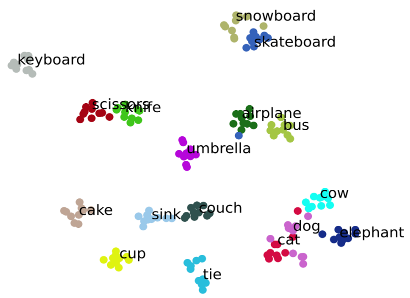

Figure 7. The CLIP's image embeddings of top retrieved boxes.

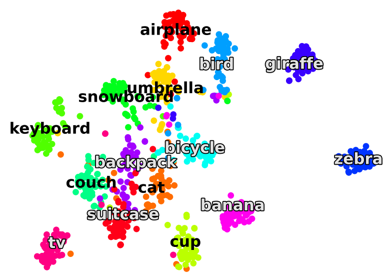

Figure 8. Embeddings of some selected **base** and novel classes.

in our case, we achieve the best performance when using $\beta = 0.8$, implying that the classification score has a greater contribution than the distillation score to the final score.

The importance of the objectness score in Eq. (6). We compare the performance of our model with and without multiplication of the objectness score o_i. The object detector's objectness score provides an indication of the presence of an object in an image. Hence, multiplying the final score by the objectness score can mitigate false positive and false negative detections. In Tab. 7, we observe a performance gain of +5.9 in AP$_{novel}$ with the multiplication of the objectness score compared to the model without it.

Reason to choose top retrieved boxes as pseudo labels. Unlike the CLIP features of random proposals, the top-retrieved boxes are distinct as visualized in Fig. 7. Therefore, these top-retrieved boxes are good candidates for training the sigmoid classifier for novel classes.

Visualization of the same embedding space of some selected base and novel classes to demonstrate the discriminative capability of our LP-OVOD is illustrated in Fig. 8.

4.3. Transfer from LVIS to Objects365 and VOC

We evaluate the transfer learning performance of our approach on Objects365 [33] and PASCAL VOC [6] datasets, following the protocol in [5, 10]. We use a pretrained model on the LVIS dataset, which includes the 'frequent' and 'common' classes, and evaluate its performance on the validation sets of Objects365 and PASCAL VOC, consisting of 365 and 20 classes, respectively. For Objects365, we use part V1 of the newly released Objects365 V2 dataset, consisting of 30,310 images and over 1.2M bounding boxes. For PASCAL VOC, we retrieve the top $K = 10$ proposals per novel class for Objects365 and the top $K = 50$ proposals for PASCAL VOC and set $\beta = 0.6$. Results are reported in Tab. 8. Our approach outperforms ViLD [10] and DetPro [5] with a substantial margin of approximately +1.5 in APs, demonstrating the effectiveness of our approach in various transfer learning settings beyond COCO and LVIS.

	Objects365			PASCAL VOC	
Method	**AP**	**AP50**	**AP75**	**AP50**	**AP75**
ViLD† [10]	10.2	16.2	10.9	72.2	56.7
DetPro [5]	10.9	17.3	11.5	74.6	57.9
Ours	**12.6**	**18.9**	**13.1**	**76.0**	**59.4**

Table 8. Transfer from LVIS to Objects365 and PASCAL VOC. †denotes the re-implementation of ViLD in the DetPro repository.

5. Discussion and Conclusion

Limitations: As shown in Tab. 1, the performance of novel classes is still lagging behind that of base classes, with a gap of 20 points in Box AP on the COCO dataset. One of the main reasons for this is that we did not fine-tune or improve the box regression for novel classes, as we only focused on the classification head. This is due to the lack of box annotations for novel classes, which is a common issue in OVOD. Additionally, CLIP's visual embeddings are not highly sensitive to the precise box location but only require that the box contains the object or important parts of the object. As a result, there is limited information available for improving the bounding boxes based solely on CLIP. Therefore, further research on improving box regression would be an interesting direction for OVOD.

Conclusion: In this work, we have introduced a simple yet effective approach for OVOD with two contributions. Firstly, we propose a linear probing approach that utilizes a pretrained Faster R-CNN to learn a highly discriminative feature representation in the penultimate layer, which is then used to train a linear classifier for novel classes. Secondly, we propose to replace the standard softmax classifier with a sigmoid classifier that is able to predict scores for each class independently, which unifies the classifier heads for both base and novel classes. Our approach outperforms strong baselines of OVOD on the COCO dataset with an AP$_{novel}$ of 40.5, setting a new state of the art.

References

[1] 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. 5

[2] M. Bravo, S. Mittal, and T. Brox. Localized vision-language matching for open-vocabulary object detection. In *German Conference on Pattern Recognition (GCPR) 2022*, 2022. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[5] 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. 2, 5, 6, 8

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 8

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 2

[8] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 5

[9] 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. *arXiv preprint arXiv:2111.09452*, 2021. 2, 5

[10] 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*, 2021. 1, 2, 3, 5, 6, 8

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 5, 7

[12] 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. 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5

[15] 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. 1, 2

[16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4

[17] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022. 5, 6

[18] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 2, 3, 5

[19] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2

[20] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 2, 5

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 4

[22] 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. 2, 5

[23] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. 2

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[25] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 2, 5

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in*

natural language processing (EMNLP), pages 1532–1543, 2014. 2

[27] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[29] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, 2020. 5

[30] 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. *arXiv preprint arXiv:2207.03482*, 2022. 2

[31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[32] 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, 3, 5, 6

[33] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 8

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[35] Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M Patel, and Ran Xu. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23539–23549, 2023. 2, 5

[36] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9919–9928. PMLR, 13–18 Jul 2020. 2

[37] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2567–2575, 2022. 2

[38] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 5

[39] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 2, 5

[40] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 2, 5

[41] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European conference on computer vision*, pages 192–210. Springer, 2020. 2

[42] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 2, 4, 5

[43] 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. 2, 5

[44] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[45] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas, et al. Exploiting unlabeled data with vision and language models for object detection. *arXiv preprint arXiv:2207.08954*, 2022. 2, 4, 5

[46] 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. 2, 5, 6

[47] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 5

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[49] 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. 5

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2