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

*The first two authors contribute equally.

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.

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.

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
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
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 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
using cosine similarity \( \cos \)

all proposals

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 \( f^\text{cls}_i \) and distillation feature \( f^\text{dis}_i \), 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 \( e^\text{image}_i \), which comes from cropped images of proposal \( \tilde{b}_i \). The distillation head is trained using L1 loss. In particular,

\[
\begin{align*}
\mathcal{L}^\text{cls}_{\text{Base}} &= \sum_i \text{Focal loss}(\text{Sigmoid}(f^\text{cls}_i; W_B), c_i), \quad (1) \\
\mathcal{L}^\text{dis} &= \sum_i \| f^\text{dis}_i - e^\text{image}_i \|_1, \quad (2)
\end{align*}
\]

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 \( e^\text{image}_i \) of all proposals \( \tilde{b}_i \) having the objectness score \( o_i \) larger than \( \tau \) in the training set. For each novel category with text embedding \( e^\text{text}_c \) 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 \( \cos(e^\text{text}_c, e^\text{image}_i) \). 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 \( f^\text{cls}_i \). The sigmoid classifier for novel classes is trained as follows:

\[
\mathcal{L}^\text{Novel} = \sum_{i=1}^{\lfloor |\mathcal{P}^+|+|\mathcal{P}^-| \rfloor} \text{Focal loss}(\text{Sigmoid}(f^\text{cls}_i; W_N), c_i), \quad (3)
\]

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 \( f^\text{cls}_i \) and distillation feature \( f^\text{dis}_i \), 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 \( s^\text{cls}_i \) is calculated as:

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

For the distillation head, we compute the distillation score \( s^\text{dis}_i \) as the softmax score of the cosine similarity between the distillation features \( f^\text{dis}_i \) and text embeddings \( e^\text{text}_c \) with temperature \( \kappa \) as:

\[
s^\text{dis}_i = \text{Softmax}_c \left( \frac{\cos(f^\text{dis}_i, e^\text{text}_c)}{\kappa} \right) \in [0, 1]^{|C_B|+|C_N|}. \quad (5)
\]

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

\[
s_i = o_i \cdot \begin{cases} s^\text{cls}_i & \text{for base classes} \\ (s^\text{cls}_i)^\beta (s^\text{dis}_i)^{1-\beta} & \text{for novel classes} \end{cases}
\]

where \( \beta \) are coefficient hyper-parameter for novel classes.
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, second best.

<table>
<thead>
<tr>
<th>Method</th>
<th>Venue</th>
<th>Training source</th>
<th>Box AP on COCO</th>
<th>Mask AP on LVIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>AP_{novel}</td>
<td>AP_{base}</td>
</tr>
<tr>
<td>OVR-CNN [45]</td>
<td>CVPR 21</td>
<td></td>
<td>22.8</td>
<td>46.0</td>
</tr>
<tr>
<td>XPM [14]</td>
<td>CVPR 22</td>
<td></td>
<td>27.0</td>
<td>46.3</td>
</tr>
<tr>
<td>RegionCLIP [46]</td>
<td>CVPR 22</td>
<td></td>
<td>31.4</td>
<td>57.1</td>
</tr>
<tr>
<td>PromtDet [8]</td>
<td>ECCV 22</td>
<td></td>
<td>26.6</td>
<td>59.1</td>
</tr>
<tr>
<td>Detic [47]</td>
<td>ECCV 22</td>
<td>instance-level labels in $C_B$ (use external datasets)</td>
<td>27.8</td>
<td>47.1</td>
</tr>
<tr>
<td>PB-OVD [9]</td>
<td>ECCV 22</td>
<td></td>
<td>30.8</td>
<td>46.1</td>
</tr>
<tr>
<td>OWL-ViT [25]</td>
<td>ECCV 22</td>
<td></td>
<td>41.8</td>
<td>49.1</td>
</tr>
<tr>
<td>VLDet [20]</td>
<td>ICLR 23</td>
<td></td>
<td>32.0</td>
<td>50.6</td>
</tr>
<tr>
<td>MS-OVIS [35]</td>
<td>CVPR 23</td>
<td></td>
<td>31.5</td>
<td>-</td>
</tr>
<tr>
<td>BARON [39]</td>
<td>CVPR 23</td>
<td></td>
<td>42.7</td>
<td>54.9</td>
</tr>
<tr>
<td>OV-DETR [42]</td>
<td>ECCV 22</td>
<td>instance-level labels in $C_B$</td>
<td>29.4</td>
<td>61.0</td>
</tr>
<tr>
<td>VL-PLM [45]</td>
<td>ECCV 22</td>
<td>known novel classes during training</td>
<td>34.4</td>
<td>60.2</td>
</tr>
<tr>
<td>ZSD [1]</td>
<td>ECCV 18</td>
<td>instance-level labels in $C_B$</td>
<td>0.31</td>
<td>29.2</td>
</tr>
<tr>
<td>PL [29]</td>
<td>AAAI 20</td>
<td>(zero-shot object detection)</td>
<td>4.12</td>
<td>35.9</td>
</tr>
<tr>
<td>DELO [49]</td>
<td>CVPR 20</td>
<td></td>
<td>3.41</td>
<td>13.8</td>
</tr>
<tr>
<td>ViLD [10]</td>
<td>ICLR 22</td>
<td></td>
<td>27.6</td>
<td>59.5</td>
</tr>
<tr>
<td>RegionCLIP\† [46]</td>
<td>CVPR 22</td>
<td></td>
<td>14.2</td>
<td>52.8</td>
</tr>
<tr>
<td>DetPro\† [5]</td>
<td>CVPR 22</td>
<td>instance-level labels in $C_B$</td>
<td>19.8</td>
<td>60.2</td>
</tr>
<tr>
<td>F-VLM [18]</td>
<td>ICLR 23</td>
<td>unknown novel classes during training</td>
<td>28.0</td>
<td>43.7</td>
</tr>
<tr>
<td>CORA [40]</td>
<td>CVPR 23</td>
<td></td>
<td>35.1</td>
<td>35.5</td>
</tr>
<tr>
<td>LP-OVOD (ours)</td>
<td>CVPR 23</td>
<td>(with OLN [17] proposals)</td>
<td>40.5</td>
<td>60.5</td>
</tr>
<tr>
<td>LP-OVOD (ours)</td>
<td>CVPR 23</td>
<td>(with OWL-ViT [25] proposals)</td>
<td>44.9</td>
<td>59.4</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Retrieval</th>
<th>Sigmoid</th>
<th>AP_{novel}</th>
<th>AP_{base}</th>
<th>AP_{r}</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>40.5</td>
<td>60.5</td>
<td>55.2</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>33.2</td>
<td>61.2</td>
<td>53.9</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>27.6</td>
<td>61.2</td>
<td>52.4</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Features</th>
<th>AP&lt;sub&gt;novel&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;base&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;AP&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>35.9</td>
<td>60.5</td>
<td>54.1</td>
</tr>
<tr>
<td>Distillation</td>
<td>19.7</td>
<td>60.5</td>
<td>49.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th># proposals</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP&lt;sub&gt;novel&lt;/sub&gt;</td>
<td>30.5</td>
<td>34.8</td>
<td>38.3</td>
<td>40.3</td>
<td>40.5</td>
<td>39.6</td>
</tr>
</tbody>
</table>

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

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<sub>novel</sub> 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<sub>novel</sub>.

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 features for classifying novel classes, we train a sigmoid classifier on top of the classification feature \(f_{cls}^{novel}\) and the distillation feature \(f_{dis}^{novel}\), 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<sub>novel</sub>, greatly outperforming that of the distillation head.

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 features for classifying novel classes, we train a sigmoid classifier on top of the classification feature \(f_{cls}^{novel}\) and the distillation feature \(f_{dis}^{novel}\), 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<sub>novel</sub>, 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 \(\beta\) 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,
### Table 7. The importance of objectness score $o_i$ in Eq. (6).

<table>
<thead>
<tr>
<th>Objectness</th>
<th>$AP_{novel}$</th>
<th>$AP_{base}$</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>40.5</td>
<td>60.5</td>
<td>55.2</td>
</tr>
</tbody>
</table>

Table 8. Transfer from LVIS to Objects365 and PASCAL VOC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Objects365 AP</th>
<th>Objects365 AP50</th>
<th>Objects365 AP75</th>
<th>PASCAL VOC AP</th>
<th>PASCAL VOC AP50</th>
<th>PASCAL VOC AP75</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLD[10]</td>
<td>10.2</td>
<td>16.2</td>
<td>10.9</td>
<td>72.2</td>
<td>56.7</td>
<td></td>
</tr>
<tr>
<td>DetPro [5]</td>
<td>10.9</td>
<td>17.3</td>
<td>11.5</td>
<td>74.6</td>
<td>57.9</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>12.6</td>
<td>18.9</td>
<td>13.1</td>
<td>76.0</td>
<td>59.4</td>
<td></td>
</tr>
</tbody>
</table>

\[1\] 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


[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


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


[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


