What Makes Good Open-Vocabulary Detector: A Disassembling Perspective

Jincheng Li*
Qihoo 360 AI Research
Beijing, China
lijincheng@360.cn

Chunyu Xie*
Qihoo 360 AI Research
Beijing, China
xiechunyu@360.cn

Xiaoyu Wu
Qihoo 360 AI Research
Beijing, China
wuxiaoyu1@360.cn

Bin Wang
Qihoo 360 AI Research
Beijing, China
wangbin10@360.cn

Dawei Leng†
Qihoo 360 AI Research
Beijing, China
lengdawei@360.cn

ABSTRACT

Open-vocabulary detection (OVD) is a new object detection paradigm, aiming to localize and recognize unseen objects defined by an unbounded vocabulary. This is challenging since traditional detectors can only learn from pre-defined categories and thus fail to detect and localize objects out of pre-defined vocabulary. To handle the challenge, OVD leverages pre-trained cross-modal VLM, such as CLIP, ALIGN, etc. Previous works mainly focus on the open vocabulary classification part, with less attention on the localization part. We argue that for a good OVD detector, both classification and localization should be parallelly studied for the novel object categories. We show in this work that improving localization as well as cross-modal classification complement each other, and compose a good OVD detector jointly. We analyze three families of OVD methods with different design emphases. We first propose a vanilla method, *i.e.*, cropping a bounding box obtained by a localizer and resizing it into the CLIP. This vanilla method totally decouples the localization and classification components, making it convenient to improve the OVD performance by applying more advanced object localization models and VLMs. However, resizing cropped regions inevitably causes the deformation of the object and leads slow calculation speed. To address these, we next introduce another approach, which combines a standard two-stage object detector with CLIP. A two-stage object detector includes a visual backbone, a region proposal network (RPN), and a region of interest (RoI) head. We decouple RPN and ROI head (DRR) and use RoIAlign to extract meaningful features. In this case, it avoids resizing objects. To further accelerate the training time and reduce the model parameters, we couple RPN and ROI head (CRR) as the third approach. We conduct extensive experiments on these three types of approaches

in different settings. On the OVD-COCO benchmark, DRR obtains the best performance and achieves 35.8 Novel AP_{50}, an absolute 2.8 gain over the previous state-of-the-art (SOTA). For OVD-LVIS, DRR surpasses the previous SOTA by 1.9 AP_{50} in rare categories. We also provide an object detection dataset called PID and provide a baseline on PID. Codes and datasets will be released.

CCS CONCEPTS

• **Computing methodologies → Object detection**.

KEYWORDS

vision-language model, open-vocabulary detection, dataset

1 INTRODUCTION

Object detection is a prominent vision task, aiming at localizing and recognizing objects in images. This task requires a variety of fine-grained annotations (*e.g.*, the bounding boxes and classes) of each object during training, which, however, makes it hard to extend the size of data since manual human annotations are costly and tedious. In this sense, traditional object detectors may fail to precisely detect and localize objects out of pre-defined vocabulary at inference.

Open-vocabulary detection (OVD), a task to detect unseen objects defined by an unbounded vocabulary, has attracted much attention in the most recent period. The core challenge of the OVD task is how to localize and classify unseen (novel) categories at the inference stage since they can only learn the knowledge from pre-defined (base) categories during training. We next analyze the corresponding existing solutions to overcome the above challenge from two perspectives: classification and localization.

To classify the novel categories, several works [1, 7, 26] leverage the excellent zero-shot generalization ability of large-scale vision-language models (VLMs) such as CLIP [14]. To this end, they modify a two-stage object detector, including using the visual encoder of CLIP as the backbone of the object detector, replacing the class-specific classification head with a class-agnostic classification head,

---

and so on. Meanwhile, they re-train or finetune VLMs for open-vocabulary detection since most VLMs are pre-trained on image-text pairs but not detection data. While detection-tailored pre-training is beneficial for the OVD task, some studies like F-VLM [11] discard this technique and also achieve significant results. In this sense, whether to use detection-tailored pre-trained CLIP remains an open question to be discussed.

On the other hand, to localize the novel categories, most existing methods [7, 20, 26] use a region proposal network (RPN), which is trained on base categories. They demonstrate that the RPN trained without seeing novel categories can generalize to localize novel categories [7]. The bounding boxes obtained by RPN play a critical role in predicting the final boxes. In this way, a better proposal network should further improve the overall performance of the OVD model since the generated bounding boxes will be more accurate. However, how to effectively improve the detection ability under the settings of the OVD task is still a challenge to be solved.

In this paper, we set out to address these issues under the settings of the open-vocabulary detection task. Our goal is to analyze which part of localization and classification can improve the overall performance of the OVD task. We use three simple but effective approaches, discuss their advantages and disadvantages and design appropriate experiments for them.

We first introduce a vanilla method for the OVD task, *i.e.*, cropping the bounding box, resizing it into the required input size of CLIP, and then feeding it into the visual encoder of CLIP. The detected bounding box is given by a pre-trained RPN, while the classification score is calculated by the cosine similarity of box embedding and the text embedding, which is extracted by the text encoder of CLIP. This vanilla method completely decouples the detection and classification components, making it easier to adopt different models. In this case, we apply different RPNs to investigate the effectiveness of RPN under the OVD settings, consisting of RPN with extra data or stronger RPN. Our goal in this experiment is to improve the detection ability, such that the overall performance of the OVD task would be further improved. In addition, we attempt to use different types of CLIP to enhance the vanilla performance. Although it is convenient to replace each component, it still suffers from two limitations: 1) it is non-trivial to extract meaningful features when facing tiny bounding boxes, *i.e.*, 2×2 pixel region. 2) it has slow inference speed due to the operations of cropping and resizing.

Next, we present another popular approach, which modifies a standard two-stage object detector [16] and combines it with CLIP. A two-stage object detector consists of a visual backbone, an RPN, and a region of interest (RoI) head. Several works [7, 11, 20] use a pre-trained RPN to extract proposals, then obtain their features via the visual encoder of CLIP relying on RoIAlign [16]. The characteristic of this approach is that it decouples RPN and ROI head (DRR) to reduce the fusion of detection features and classification features. Compared to the vanilla method, DRR is able to extract significant features for tiny objects and has a faster inference speed. Unfortunately, it applies two backbones to decouple RPN and ROI head, leading to require more computational costs and training time.

Based on DRR, we propose the third approach, which couples RPN and ROI head (CRR). CRR uses one backbone for localization and classification. In this sense, it accelerates the training time

and reduces the parameters of the model. To summarize, our main contributions are as follows:

- We investigate and analyze the advantages and disadvantages of three approaches for open-vocabulary detection. We demonstrate that both localization and classification can improve open-vocabulary detectors.
- We design appropriate experiments for three approaches with the commonly used techniques and achieve state-of-the-art results on OVD-COCO and OVD-LVIS benchmarks.
- We propose a product dataset (PID) with human annotations for the open-vocabulary detection task and provide a strong baseline on PID.

2 RELATED WORK

Vision-Language Pre-training. Vision-language pre-training aims to learn the correspondence between vision and natural language. It is attractive that pre-trained vision-language models (VLMs) [9, 14, 22, 23] trained on large-scale image-text pairs show excellent zero-shot/few-shot migration ability on classification, object detection, and instance segmentation tasks. In particular, F-VLM [11] shows that pre-trained VLMs have a strong generalization ability when transferring to the OVD task. Meanwhile, F-VLM eliminates the need for knowledge distillation or detection-tailored pre-training. This motivates us to investigate how to make better use of VLMs on the open-vocabulary object task. In this paper, we apply different VLMs in the fundamental approaches, including original VLMs, larger VLMs, and detection-tailored pre-training VLMs.

Open-Vocabulary Detection. Traditional object detection may fail to localize and recognize unseen objects in an image at inference, *i.e.*, zero-shot object detection. Recently, OVR-RCNN [24] proposes the open-vocabulary detection (OVD) benchmark to detect and localize objects for which no bounding box annotation is provided during training. While OVR-RCNN evaluates the models on tens of categories, ViLD [7] proposes to evaluate on more than 1,000 categories, *i.e.*, LVIS [8]. Following the OVD benchmark, most existing methods are proposed with different forms of weak supervision, such as extra image-caption pairs [6, 21, 24, 26], extra classification datasets [28], and vision-language pre-trained models [3, 5, 7, 25] like CLIP [14]. For example, RegionCLIP [26] introduces a region-level pre-training alignment method with extra image-text pairs *e.g.*, CC3M [18], demonstrating its capability on zero-shot and OVD task transfer learning. Detic [28] trains the classifiers of a detector on image classification data, *i.e.*, ImageNet-21K [17], yielding excellent detectors even for classes without box annotations. BARON[20] proposes to align the embedding of the bag of regions beyond individual regions, relying on the generalization ability of large-scale vision-language pre-trained models. These methods keep achieving better results than previous state-of-the-art methods with various techniques, such as more powerful offline proposal generators [19], knowledge distillation [13], and prompt learning [4]. When facing a new dataset for new scenarios, it leaves researchers and engineers confused about which approaches or techniques to use. In this paper, we summarize three fundamental approaches for open-vocabulary detection and investigate them with different techniques, showing surprising results of different combinations.
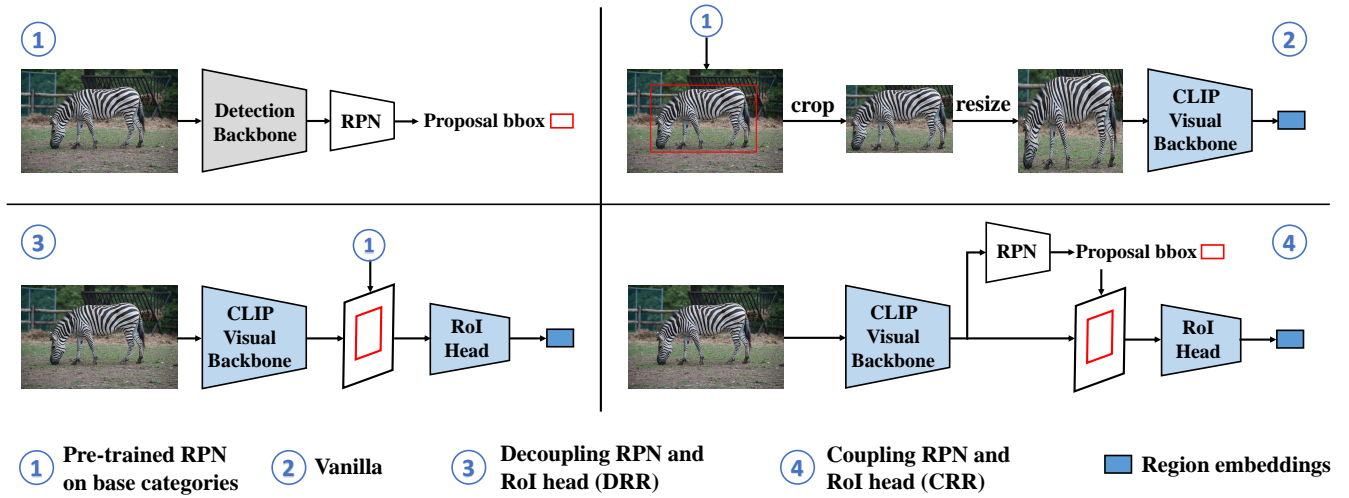
Figure 1: An overview of three approaches of open-vocabulary detection: a vanilla method, decoupling RPN and ROI head (DRR), and coupling RPN and ROI head (CRR). Note that Vanilla and DRR require two backbones while CRR only needs one visual backbone. For simplicity, we omit the logits of regions when training RPN on base categories. For all three fundamental approaches, we get the classification scores via cosine similarity, calculated by the region embeddings and the text embeddings, where the text embeddings are obtained by the text encoder of CLIP.

3 APPROACH

In this paper, we introduce three approaches for open-vocabulary detection: a vanilla method, decoupling RPN and ROI head (DRR), and coupling RPN and ROI head (CRR). The overview of the approaches is illustrated in Figure 1.

3.1 A Vanilla Method

To solve the open-vocabulary detection problem, we attempt to isolate the open-vocabulary detection task into two independent sub-tasks, that is, object localization and object classification.

Object Localization. To localize objects, both the class-aware object detector, such as YOLO [15], Faster R-CNN , and the class-agnostic localizer, such as RPN, OLN[10] can be used. The first challenge for OVD task is to localize novel objects. To solve this problem, we adopt the detection backbone with RPN to localize all objects and classify them as the foreground class. Furthermore, we improve RPN using Faster R-CNN. Here, the multi-classes head within Faster R-CNN is changed to a class-agnostic head so that it can resolve a binary task like RPN. The class-agnostic module within RPN or modified Faster R-CNN can generalize to novel objects [7]. Besides, we can use a classification-free network such as OLN [10] for the purpose of localizing objects. In this case, we can modify the training loss to estimate the objectness of each region purely since localization-related metric tends to be robust to novel objects in the open world [10].

Object Classification. After localizing object candidates, we leverage a pre-trained large-scale vision-language model (VLM) such as CLIP to classify them. Specifically, We crop and resize the object candidates, and feed them into the visual encoder of VLM to achieve the corresponding region embeddings. In order to provide more context cues, these region embeddings are ensemble from 1× crop (crop the image according to the bounding box) and 1.5× crop

(extend the crop size to 1.5 times). We feed the category texts with a set of prompt templates and then feed them into the text encoder of VLM to obtain the text embeddings. Finally, The cosine similarity is calculated by the averaged text embeddings and region embeddings, and the per-class NMS is adopted to obtain the final detection results.

As this solution decomposes the OVD task into the object localization and classification sub-tasks completely, it can be easily assembled by different detectors and VLMs, leading to a powerful performance. In the vanilla method, each cropped object region needs to be fixed to the same scale for the VLMs. On the one hand, these operations reduce the efficiency of the model. On the other hand, the deformation of the regions brings more difficulties to object detection, especially for small targets.

3.2 Decoupling RPN and ROI head (DRR)

A two-stage object detector consists of a visual encoder backbone, a region proposal network (RPN), and a region of interest (RoI) head. Here, the RPN is trained end-to-end to generate high-quality region proposals and then provided for detection in the ROI head. While the traditional two-stage object detector jointly trains RPN and the ROI head, several works [19, 26] propose to decouple them to avoid the conflict that the sensitivity of the classification head to novel categories hampers the universality ability of the RPN head. So far, there are no detailed discussions about the effect of whether to decouple RPN and ROI head or not in existing works. In this paper, we try to analyze them and hope to provide a view of the balance of model performance and cost for researchers.

As aforementioned, we first introduce a training approach that decouples the RPN and ROI head in a two-stage object detector. Following existing works [1, 19, 26], we use different backbones for localization and classification. That is, one backbone is designed for RPN and the other is for ROI head. Given an image, we first use a

pre-trained RPN to obtain a set of regions of interest (*i.e.*, proposals). We then use the visual encoder (*e.g.*, ResNet-50) of CLIP [14] to encode the whole image. Given proposal boxes of an image, we extract their features along the first 3 blocks based on the whole image encoding and pool them using RoIAlign. The pooled features are then encoded by the last block of the visual encoder. We also replace the class head of Fast R-CNN with the text embeddings encoded by the CLIP text encoder. Considering the setup of open-vocabulary detection [24], we use the base class and hand-crafted prompt as the input of the text encoder of CLIP during training. During inference, we replace the base class with a combination of base and novel classes to generate new text embeddings.

3.3 Coupling RPN and ROI head (CRR)

Decoupling proposal generation and ROI head is an efficacious method to keep the universality ability of the proposal generation stage since the proposal generation stage has a class-agnostic classification, which can be easily extended to novel classes. However, this decoupling scheme means that RPN and ROI modules use different backbone encoders, which increases the training cost. In the stage of model deployment and inference, this solution will also bring a lot of extra computing time. Considering the computational efficiency, we practice the scheme of sharing the visual backbone and conduct detailed experiments. These experiments can provide researchers with speed-accuracy trade-offs.

4 EXPERIMENTS

Dataset and Metrics. We comprehensively evaluate three fundamental approaches of OVD task on COCO [12], LVIS [8], and our proposed Product Image Dataset (PID) benchmarks. COCO is a standard dataset comprising 80 categories of common objects in a natural context. It contains 118k images with bounding boxes and instance segmentation annotations. We follow OVR-CNN [24] to split the object categories into 48 base categories and 17 novel categories. We also follow ViLD [7] for the LVIS dataset to split the 337 rare categories into novel categories and the rest common and frequent categories into base categories. For simplicity, we denote the open-vocabulary benchmarks based on COCO and LVIS as OVD-COCO and OVD-LVIS, respectively. Following ViLD [7], the Novel AP_{50} and AP_r are the main metric on OVD-COCO and OVD-LVIS, respectively. Our open-vocabulary object detectors are trained on base classes. Besides, we split PID into the base and novel categories. The detailed introductions can refer to Sec 5.1.

The dataset used for detector pre-training is established from the BigDetection dataset [2]. As the BigDetection dataset has nearly covered all categories in COCO, we remove the COCO images and delete both novel and base categories of the COCO dataset from the annotations. Some categories with less than 100 train samples are removed as well. We also exclude LVIS data from the BigDetection dataset. Finally, we establish this new BigDetection dataset named BigDetection* (BD*) which has 489 categories for detector pre-training.

Implementation Details of the vanilla method. In the OVD-COCO and OVD-LVIS experiments, We use Faster R-CNN as the object detector and frozen CLIP as the object classifier. SGD optimizer is adopted for the 8× training schedule. We train the detector

for 720k iterations and divide the learning rate by 10 at 660k and 700k iterations. We also adopt linear warmup for the first 1,000 iterations starting from a learning rate of 0 to 0.02. Then, we fine-tune the model on base categories of OVD-COCO and OVD-LVIS separately. For OVD-COCO, we train 90k iterations and scale down the learning rate at 60k and 80k. For OVD-LVIS, we train 180k iterations and scale down at 120k and 160k. For these two datasets, the learning rate increases to 0.0002 for the first 5k iterations for the warmup.

In the PID experiments, we use Faster R-CNN pre-trained on the BigDetection dataset [2] as the object detector. ProductCLIP with prompt tuning, which is introduced in Sec 5.3, is applied as the object classifier. The detectors are fine-tuned with the 1x training schedule. A warm-up step with a learning rate of 0.001 is performed for the first 400 iterations. On both the pre-training and fine-tuning stages, we select 16 samples per GPU with the class-aware sampler. Multi-scale training is adopted with the short edge in the range [640, 800] and the long edge up to 1333. We use 8 A100 GPUS to perform the experiments.

Implementation Details of DRR and CRR. In Figure 1, the detection backbone and the CLIP visual backbone are both ResNet50. We adopt ImageNet [17] and RegionCLIP [26] pre-trained parameters for the detection backbone and the CLIP visual backbone, respectively. During training, we first train the RPN to obtain the proposal bounding boxes on the base categories. We then use a Faster R-CNN with ResNet50-C4 architecture as the detector and 1x training schedule (90k iterations). We train the detector using the offline pre-trained RPN. That is, we decouple the RPN and the ROI head. We use an SGD optimizer with a learning rate of 0.002 and a min-batch of 16. We set the weight of the background category to 0.2 and 0.8 for OVD-COCO and OVD-LVIS, respectively. We use the same learning rate and input-scale strategy as the vanilla method. We replace the base categories in the classification head with base + novel categories during inference. We use the top-ranked 100 proposals at test time for all detectors. For CRR, we keep the same settings as DRR, except that CRR removes the detection backbone, as shown in Figure 1.

We use the same settings as OVD-COCO when conducting experiments on PID except that the pre-trained weight of CLIP visual backbone. As aforementioned in the vanilla method, we use Product-CLIP as the pre-trained weight on PID.

4.1 Comparison with State-of-the-arts.

4.1.1 Experiment on OVD-COCO. We compare all three kinds of basic methods with most existing state-of-the-art methods on OVD-COCO. Table 1 summarizes the results. Although the vanilla method is flexible to replace the detector and classification model, it achieves comparable results on novel categories but obtains bad results on base categories. One possible reason is that CLIP has gaps between images and resized images. Moreover, CRR obtains a higher Novel AP_{50} than RegionCLIP, but lower than BARON. However, both RegionCLIP and BARON are with extra backbone, which brings more computational complexity. DRR achieves the best results and outperforms BARON by 2.7 Novel AP_{50}. These results demonstrate that DRR has the potential to be the best baseline when selecting approaches for open-vocabulary detection.

Table 1: Performance on OVD-COCO compared with state-of-the-art methods.

Method	Extra Dataset	Backbone	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
OVR-CNN [24]	COCO Captions	ResNet50	22.8	46.0	39.9
ViLD [7]	-	ResNet50	27.6	59.5	51.2
Detic [28]	COCO Captions	ResNet50	27.8	47.1	45.0
RegionCLIP [26]	CC3M	ResNet50	31.4	57.1	50.4
BARON [20]	COCO Captions	ResNet50	33.1	54.8	49.1
Vanilla (Ours)	-	ResNet50	31.8	37.2	35.5
CRR (Ours)	CC3M	ResNet50	32.0	52.5	47.1
DRR (Ours)	CC3M	ResNet50	**35.8**	54.6	49.6

Table 2: Performance on OVD-LVIS compared with state-of-the-art methods.

Method	Backbone	Require Novel Class [20]	AP_r	AP_c	AP_f	mAP
RegionCLIP [26]	ResNet50	✗	17.1	27.4	34.0	28.2
ViLD [7]	ResNet50	✓	16.7	26.5	34.2	27.8
BARON [20]	ResNet50	✓	20.1	28.4	32.2	28.4
Vanilla (Ours)	ResNet50	✗	17.2	14.8	11.5	13.9
CRR (Ours)	ResNet50	✗	14.0	23.7	28.5	21.9
DRR (Ours)	ResNet50	✗	20.1	29.9	35.7	30.5
DRR (Ours)	ResNet50	✓	**22.0**	25.4	33.7	28.1

Table 3: Influence of object localization on OVD-COCO. Faster R-CNN* denotes that it is pre-trained on BigDetection*.

Method	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
RPN	16.6	20.3	19.2
Faster R-CNN	19.5	35.7	30.9
OLN	26.2	27.0	26.6
Faster R-CNN*	29.6	32.1	31.1

Table 4: Influence of object classification on OVD-COCO.

Classifier	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
CLIP-RN50	26.6	27.4	26.8
CLIP-ViT-B	29.6	32.1	31.1
CLIP-ViT-L	31.8	37.2	35.6

Table 5: Effect of image embedding ensemble on OVD-COCO.

Method	Ensemble	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
Vanilla	✗	27.3	28.9	28.5
Vanilla	✓	29.6	32.1	31.1

4.1.2 Experiment on OVD-LVIS. We further conduct experiments and compare them with state-of-the-art methods on a larger open-vocabulary dataset, *i.e.*, OVD-LVIS. We follow the setup of RegionCLIP [26] and provide the results in Table 2. For a fair comparison with the previous SOTA (*i.e.*, BARON [20]), we also report the ensemble results which require novel class following BARON. DRR achieves 20.1 AP_r, which is significantly better than Region-CLIP by 3.0 AP_r. Meanwhile, DRR becomes the new SOTA on OVD-LVIS with the setting of requiring novel class during inference. Similar to OVD-COCO, CRR still leads a competitive result. Instead, the vanilla method obtains bad results compared to other methods, indicating that the operation of crop and resize is non-trivial to recognizing small objects in OVD-LVIS.

4.2 Analysis of Three Fundamental Approaches.

To investigate the behavior of different fundamental approaches of OVD task, we conduct several ablation studies. We design different experiments according to their intrinsic characteristics.

4.2.1 Vanilla. As this two-stage framework isolates the object localization and classification completely, it can be easily assembled by different models. With the help of advanced object detection models and vision-language models, OVD can be improved flexibly by applying different object localization and object classification modules.

The influence of object localization. Table 3 shows the influence of different object localizers. All the experiments use CLIP-ViT-B/32 [14] as the classifier. We attempt several different object localization networks including RPN, Faster R-CNN, and OLN [10] with different training schedules. We evaluate the models on the OVD-COCO dataset. As Faster R-CNN can achieve more accurate bounding boxes than RPN, it leads to a higher AP50 on both novel and base categories. Besides, in order to improve the performance

Table 6: Comparisons of different RPNs over ResNet50 on OVD-COCO. Here, COCO-48-RPN, BD*-489-RPN, and BD*-489-COCO-48-RPN denote that the RPNs are trained on base categories of OVD-COCO, BigDetection, and the combination of base categories of OVD-COCO and BigDetection, respectively.

Method	Proposal Generation Type	Pre-trained RPN	Multiplying RPN Score	Novel AP$_{50}$	Base AP$_{50}$	Overall AP$_{50}$
DRR	RPN	COCO-48-RPN	✗	31.0	56.7	50.0
DRR	RPN	COCO-48-RPN	✓	31.1	53.9	48.0
DRR	Faster R-CNN (class-agnostic)	BD*-489-RPN	✗	30.5	52.1	46.5
DRR	Faster R-CNN (class-agnostic)	BD*-489-RPN	✓	32.3	40.4	38.3
DRR	Faster R-CNN (class-agnostic)	BD*-489-COCO-48-RPN	✗	30.5	55.6	49.0
DRR	Faster R-CNN (class-agnostic)	BD*-489-COCO-48-RPN	✓	**35.8**	54.6	49.4

Table 7: Effect of CLIP visual backbone on OVD-COCO compared with state-of-the-art methods.

Method	Visual Backbone	Detection-tailored Pre-training	Novel AP$_{50}$	Base AP$_{50}$	Overall AP$_{50}$
RegionCLIP [26]	ResNet50	✗	14.2	52.8	42.7
RegionCLIP [26]	ResNet50	✓	31.4	57.1	50.4
DRR (Ours)	ResNet50	✓	**35.8**	54.6	49.4
RegionCLIP [26]	ResNet50x4	✓	39.3	61.6	55.7
DRR (Ours)	ResNet50x4	✓	**41.9**	57.8	53.7

for the novel class, we also evaluate the class-agnostic object localization network like OLN. OLN learns generalizable objectness and tends to propose any objects in the image. Compared with the Faster R-CNN, it outperforms 6.7% AP50 on novel categories. To obtain better generalization ability, we pre-trained the Faster R-CNN on the BigDetection* dataset. From the fourth row of Table 3, we observe that the additional pre-training brings a 3.4% AP50 improvement.

The influence of object classification. Table 4 shows the influence of different classifiers. According to the above experiments, we use the Faster R-CNN pre-trained on the BigDetection* dataset for object localization. CLIP [14] models are used to classify cropped region proposals. Compared with CLIP-ResNet50 and CLIP-ViT-B/32, CLIP-ViT-L/14 shows better performance. This experiment proves that a strong open-vocabulary object classification model can provide the powerful capability for detecting novel objects.

The influence of image embedding ensemble. After localizing objects, we crop and resize the object regions for the object classifier to compute image embedding. There are limited contextual cues if cropping directly according to the bounding box localized by the object detector. We attempt to expand the bounding boxes by 1.5 times and ensemble the image embeddings from 1× crop and 1.5× crop. We use the Faster R-CNN pre-trained on the BigDetection* dataset for object localization and CLIP-ViT-B/32 for object classification. Table 5 shows that the image embedding ensemble can improve the OVD performance efficiently.

4.2.2 DRR. **The effect of RPN.** As illustrated in Figure 1, we first use an offline RPN to obtain proposal bounding boxes and then extract the corresponding features with the help of the visual encoder of CLIP. In this section, we investigate the effect of different RPN under the settings of DRR. We train several RPNs on different datasets to determine whether a proposal is a foreground. We also replace RPN with Faster R-CNN with a class-agnostic head to improve the

detection performance. Compared to RPN, Faster R-CNN prefer to output more accurate bounding boxes with more high-quality objectness logits, thus we attempt to multiply the logits with the final CLIP scores, called "Multiplying RPN score". That is, it can be formulated as $\sqrt{s_1 \cdot s_2}$, where s_1 and s_2 denote the objectness logits and the CLIP score, respectively. Table 6 summarizes the results.

From Table 6, we get the following observations. 1) Replacing RPN with Faster R-CNN cannot achieve the expected results. Generally, we can improve the overall model performance by improving the offline RPN. However, COCO-48-RPN and BD*-489-COCO-48-RPN obtain similar Novel AP$_{50}$ when without multiplying RPN score (31.0 vs. 30.5), indicating a better offline RPN does not work as well. This is because the offline RPN only provides bounding boxes for the visual backbone of CLIP and does not directly participate in the loss of classification. 2) The significant objectness logits within a better offline RPN are indeed important for model performance. We find that DRR with "multiplying RPN score" leads to a Novel AP$_{50}$ of 35.8 compared to the one without "multiplying RPN score", which becomes a new SOTA over ResNet50 on OVD-COCO.

The effect of CLIP visual backbones. From Table 7, we argue that the detection-tailored pre-training like RegionCLIP [26] is required when needing the best performance. We then use a larger visual encoder to extract the features of regions of interest. We use ResNet50x4 as the backbone and conduct experiments on OVD-COCO. Relying on the observation from Table 6, we use BD*-489-COCO-48-RPN as the RPN and multiply the CLIP scores with RPN scores during inference for DRR. In Table 7, DRR surpasses the previous state-of-the-art (*i.e.*, RegionCLIP [26]) by 2.6 AP$_{50}$ in novel categories, which benefits from more accurate bounding boxes and more significant objectness logits. Meanwhile, ResNet50x4 shows better results than ResNet50 from Tables 7 & 1. This implies that a model with a larger capacity is able to obtain better representations and improve overall performance.
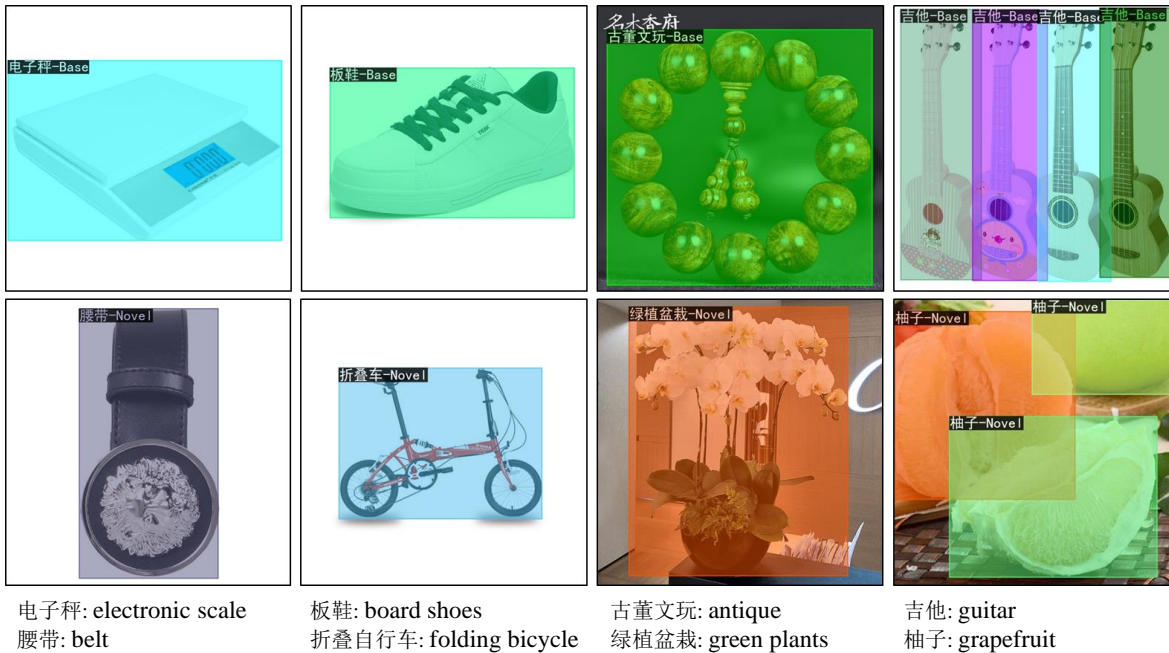
电子秤: electronic scale
腰带: belt

板鞋: board shoes
折叠自行车: folding bicycle

古董文玩: antique
绿植盆栽: green plants

吉他: guitar
柚子: grapefruit

Figure 2: Examples (with annotations) of PID. The first and second rows are from the base and novel categories, respectively.

Table 8: Comparisons of the computational efficiency over ResNet50 on OVD-COCO.

Method	Params ↓	FPS (BS1 A100) ↑
Vanilla	136.9 M	2
DRR	143.4 M	12
CRR	**111.6 M**	**13**

Table 9: Comparisons of different fundamental approaches over ResNet50 on PID. *More analysis can be found in Section 5.2.

Method	Visual Backbone	Generalized (233+233)		
		Novel	Base	Overall
Vanilla*	ResNet50	**42.0**	52.8	47.4
DRR	ResNet50	30.7	35.6	33.2
CRR	ResNet50	27.6	34.3	31.0

4.2.3 CRR. **Analysis of Computational efficiency.** Table 8 shows the performance and computational efficiency of the three frameworks. For a fair comparison, we obtain the results of different models under the same environment. We use the Detectron2 [1] toolbox based on PyTorch [2] and an A100 GPU to calculate the FPS. The batch size (denoted as BS in Table) is set to 1. From Table 8, we can see that the scheme of sharing the visual backbone has the highest computational efficiency compared to the other two models, and it is far more accurate than the scheme of CLIP on cropped regions. Sharing the visual backbone is indeed more effective in specific real-world scenarios.

5 PRODUCT IMAGE DATASET–A NEW OBJECT DETECTION DATASET

5.1 Introduction of PID.

Product Image Dataset (PID) is an object detection dataset consisting of a series of product images along with the corresponding human-annotated bounding boxes and classes. Note that we label all

bounding boxes in Chinese. PID contains 37,540 images and 52,796 annotations. The image size of PID is 800×800. To adapt to the setting of open vocabulary object detection like OVR-RCNN [24], we split PID into a training dataset and a test dataset, and put more classes in the test dataset. Here, the training dataset consists of 233 classes, 14,802 images, and 20,690 annotations while the test dataset includes 466 classes, 22,738 images, and 32,106 annotations. That is, we use 233 classes as base categories and 233 classes as novel categories. We provide several examples with annotations in Figure 2. The product image dataset will be released.

5.2 Finetuned results on PID.

We train three approaches on the base categories of PID and then evaluate them on the base and novel categories. We adopt ResNet50 as the visual backbone of CLIP. The results are shown in Table 9. Somewhat surprisingly, the vanilla method achieves the best performance compared to DRR and CRR, which is inconsistent with the results on OVD-COCO and OVD-LVIS. One possible reason is that the objects in PID are large such that cropping and resizing the proposal is an effective way to improve the overall performance.

Table 10: Effect of RPNs and the ensemble score on PID.

Visual Encoder Pre-training	Pre-trained RPN	Multiplying RPN Score	Generalized (233+233)		
			Novel	Base	Overall
ProductCLIP	PID-233-RPN	✗	22.0	34.1	28.1
ProductCLIP	PID-233-RPN	✓	26.6	36.2	31.4
ProductCLIP	BD*-489-PID-233-RPN	✓	30.7	35.6	33.2

Table 11: Effect of knowledge distillation on PID. We use L1 distillation by default.

Visual Encoder Pre-training	Pre-trained RPN	Temperature	Generalized (233+233)		
			Novel	Base	Overall
ProductCLIP	PID-233-RPN	-	22.0	34.1	28.1
ProductCLIP	PID-233-RPN	1	22.6	34.5	28.6
ProductCLIP	PID-233-RPN	10	21.1	31.7	26.4
ProductCLIP	PID-233-RPN	100	16.3	24.5	20.4

Table 12: Effect of the number of prompts of the text encoder on PID.

Visual Encoder Pre-training	Pre-trained RPN	#prompt	Generalized (233+233)		
			Novel	Base	Overall
ProductCLIP	PID-233-RPN	93	22.0	34.1	28.1
ProductCLIP	PID-233-RPN	1	22.3	34.3	28.3

Table 13: Effect of prompt tuning on PID.

Visual Encoder Pre-training	Pre-trained RPN	Prompt Tuning	Generalized (233+233)		
			Novel	Base	Overall
ProductCLIP	BD*-489-PID-233-RPN	✗	27.1	35.8	31.5
ProductCLIP	BD*-489-PID-233-RPN	✓	28.1	35.6	31.9

5.3 Ablation study on PID.

Summary of ablation study. Several related works [1, 7, 20, 26] in open-vocabulary detection have applied a variety of techniques to improve the model performance. However, it is worth noting that some works use a part of these techniques while others do not, which makes this line of work confusing. That is, it is hard to use existing techniques to improve the model performance when transferring to a new dataset, *e.g.*, PID. Here, we investigate the effectiveness of commonly used techniques on PID, including different proposal generators, multiplying RPN score, knowledge distillation, and prompt learning. We use DRR to conduct all experiments on PID.

Why use DRR to conduct ablation study on PID? From Table 9, it seems that the vanilla method is a better choice compared to CRR and DRR. However, it suffers from slow inference speed [7] and achieves unsatisfactory results on the two popular public OVD datasets, *i.e.*, OVD-COCO and OVD-LVIS (see Tables 1 & 2 & 9). Instead, DRR outperforms the vanilla method and CRR on OVD-COCO and OVD-LVIS and is with excellent computational

efficiency (see Table 8). Since the main goal is to investigate which technique is more effective when transferring to new datasets, we adopt DRR to conduct the ablation study on PID.

The effect of RPNs and the ensemble score. First, we train a CLIP model using image-text product pairs and name it ProductCLIP. Note that the text encoder of ProductCLIP is designed for Chinese. We then train an offline RPN on the base categories of PID for DRR, called PID-233-RPN. Similar to Table 6, we investigate the effectiveness of "multiplying RPN score" during inference. We also train a BD*-489-PID-233-RPN to further improve the quality of bounding boxes and objectness logits. From Table 10, the model with "multiplying RPN score" can achieve better detection performance over base and novel categories. When improving the performance of the offline RPN, the Novel AP_{50} improves from 26.6 to 30.7. We can obtain similar observations as in Table 6.

The effect of knowledge distillation. Knowledge distillation is a commonly used technique to bridge the gap between the object-level representations from detectors and the image-level representations from CLIP. Thus, they can obtain significant classification scores

from the object-level representations and the embeddings from the text encoder of CLIP. Following [1, 7], we use L1 knowledge distillation to check whether this technique is valid for PID. We also adjust different values of temperature in knowledge distillation to find a good hyperparameter. Table 11 summarizes the results. We observe that DRR with knowledge distillation (KD) achieves slightly better results than the one without KD, *i.e.*, 22.6 vs. 22.0. One possible reason is that we pre-train the visual backbone of CLIP using image-text product pairs, that is, there is no gap between the object-level representations from the visual backbone with the text encoder.

The effect of prompt numbers. In the setup of open-vocabulary detection, we usually obtain the text embeddings by directly feeding the concepts with human-crafted prompts into the text encoder of CLIP. Here, We evaluate different number of prompts of the text encoder in Table 12. We observe that the effectiveness of the number of prompts can be negligible on PID.

The effect of prompt tuning. Next, we introduce another common technique in prompt modification. We replace the human-crafted prompt with a learned prompt by prompt tuning [27]. Table 13 summarizes the results. The models with prompt tuning can improve by 1.0 Novel AP_{50}, indicating that prompt tuning is an effective technique to improve the model performance.

6 CONCLUSION AND SOCIAL IMPACTS

In this paper, we have conducted a comprehensive study of three fundamental commonly used approaches for open-vocabulary detection, a vanilla method, decoupling RPN and ROI head (DRR), and coupling RPN and ROI head (CRR). We analyze the advantages and the disadvantages of the three approaches and give a discussion about how and when to use them. Extensive experiments demonstrate the effectiveness of different approaches under different setups on COCO and LVIS benchmarks. Besides, we propose a product dataset for object detection called PID and provide a strong baseline on PID. We hope our analyses and datasets promote the development of open-vocabulary detection.

REFERENCES

[1] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems* 35 (2022).

[2] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. 2022. BigDetection: A Large-scale Benchmark for Improved Object Detector Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4777–4787.

[3] Peixian Chen, Kekai Sheng, Mengdan Zhang, Yunhang Shen, Ke Li, and Chunhua Shen. 2022. Open Vocabulary Object Detection with Proposal Mining and Prediction Equalization. *arXiv preprint arXiv:2206.11134* (2022).

[4] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022. 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*. 14084–14093.

[5] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. 2022. Promptdet: Expand your detector vocabulary with uncurated images. In *European Conference on Computer Vision*.

[6] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*. 266–282.

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5356–5364.

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. 4904–4916.

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

[11] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. 2023. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. In *International Conference on Learning Representations*.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 740–755.

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

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

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 779–788.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115 (2015), 211–252.

[18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[19] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. 2023. Detecting Everything in the Open World: Towards Universal Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

[21] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. 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*.

[22] Chunyu Xie, Heng Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. 2022. Zero and R2D2: A Large-scale Chinese Cross-modal Benchmark and A Vision-Language Framework. *arXiv preprint arXiv:2205.03860* (2022).

[23] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335* (2022).

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

[25] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. 2022. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*. 159–175.

[26] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16793–16803.

[27] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

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