This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Exploring Region-Word Alignment in Built-in Detector for Open-Vocabulary Object Detection

Heng Zhang[1]* Qiuyu Zhao[1]* Linyu Zheng[1]† Hao Zeng[1]
Zhiwei Ge[1] Tianhao Li[1] Sulong Xu[1]
[1] JD.com
{zhangheng291, zhaoqiuyu3, zhenglinyu1, zenghao30, gezhiwei, litianhao5, xusulong}@jd.com

Abstract

Open-vocabulary object detection aims to detect novel categories that are independent from the base categories used during training. Most modern methods adhere to the paradigm of learning vision-language space from a large-scale multi-modal corpus and subsequently transferring the acquired knowledge to off-the-shelf detectors like Faster-RCNN. However, information attenuation or destruction may occur during the process of knowledge transfer due to the domain gap, hampering the generalization ability on novel categories. To mitigate this predicament, in this paper, we present a novel framework named BIND, standing for Bulit-IN Detector, to eliminate the need for module replacement or knowledge transfer to off-the-shelf detectors. Specifically, we design a two-stage training framework with an Encoder-Decoder structure. In the first stage, an image-text dual encoder is trained to learn region-word alignment from a corpus of image-text pairs. In the second stage, a DETR-style decoder is trained to perform detection on annotated object detection datasets. In contrast to conventional manually designed non-adaptive anchors, which generate numerous redundant proposals, we develop an anchor proposal network that generates anchor proposals with high likelihood based on candidates adaptively, thereby substantially improving detection efficiency. Experimental results on two public benchmarks, COCO and LVIS, demonstrate that our method stands as a state-of-the-art approach for open-vocabulary object detection.

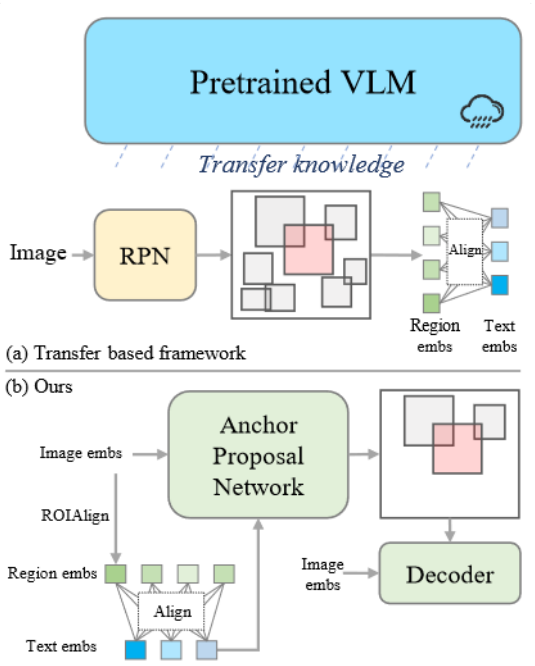

Figure 1. Comparison with the transfer-based framework and our built-in detector framework. **(a)** Transfer-based methods learn from VLM and information attenuation may occur during the process of knowledge transfer. **(b)** Our built-in detector learns region-word alignment directly where an anchor proposal network is developed for region localization, *i.e.*, reliable positional recommendations for efficient training and inference.

1. Introduction

Object detection is a fundamental task in computer vision that involves object recognition and localization in images [16]. Traditional object detection methods, such as RCNN [7], Faster-RCNN [24] and DETR [1], are primarily designed for close set scenarios, where the target categories to be detected remain consistent throughout both the training and inference phases. When applying them in real-world scenarios, the challenge arises of detecting novel classes that have not been seen during the training phase. To tackle this issue, a new paradigm, open vocabulary object detection (OVD) [28, 32], has been proposed recently and received much attention in the research community. The core of this paradigm is to acquire semantic understand-

*Equal contributions.
†Corresponding author.

ing in visual-language space through large-scale image-text pairs, effectively serving as a bridge to extend traditional detection models into the realm of OVD.

Many modern OVD methods [28] employ CLIP [22] as the visual-language model (VLM) for its simplicity and strong zero-shot generalization ability, and transfer the vision-language alignment ability from CLIP to detection models for open-set object classification and localization, as shown in Figure 1(a). However, the direct utilization of CLIP is less compatible with OVD due to its requirements for region-word alignment [8], given that CLIP is trained to globally align an image with its associated text. The following two facts confirm this issue: (i) For two regions containing the same object, CLIP tends to assign a higher score to the one with more backgrounds [34]. (ii) For classification tasks, CLIP obtains 60% top-1 accuracy on ImageNet [25] with whole classification, but only 19% on LVIS [9] with region classification. To alleviate this problem, a series of studies [18, 34] develop fine-tuning and/or adaptation-based methods. Despite these efforts, the limitations of pre-trained models hinder their ability to achieve satisfactory performance. Another series of studies [27, 29, 30] exploit region-word pre-training and/or distillation-based methods. However, they depend on external detectors with manually designed anchors which lack adaptive capabilities, resulting in excessive computational redundancy and suboptimal performance.

To alleviate the above dilemmas, in this paper, we propose a novel architecture that can not only learn region-word alignments from large-scale image-text pairs but also directly apply the learned knowledge to the built-in detectors. Specifically, as shown in Figure 2, we adopt an Encoder-Decoder structure and design the following two training stages.

1) Region-word alignment. In this stage, the goal is to learn a fine-grained visual-language alignment space with large-scale image-text pairs by region-word matching. Given an input image-text pair, an RPN is employed to obtain a set of regions in the image and the corresponding text description is segmented into a set of words. Then, the region-word alignment can be formulated as a bipartite matching problem between image regions and word candidates which can be solved by the Hungarian matching algorithm. Finally, the focal loss [17] is calculated. Especially, we found that the sensitivity of the trained visual-language model to image regions is important in our method. Therefore, we employ DINO-v2 [20] as the image encoder to extract the backbone features of input images, which is able to generate segment-level feature maps of instances and is proven beneficial for learning region-word alignment and locating objects.

2) Built-in detector training. In this stage, the goal is to equip the model with the ability to locate and classify objects. Here, a DETR-style [1] decoder is trained which receives the vision embeddings and language embeddings output from the previous stage for predicting the target bounding box and class.

In this way, as shown in Figure 1(b), the proposed method does not require knowledge transfer from the pre-trained VLMs to off-the-shelf detectors, alleviating the overfitting to base classes and improving the generalization ability to novel classes. In addition, we also propose an anchor proposal network that can adaptively provide simplified proposals based on images and text queries, significantly reducing proposal redundancy and computations.

The main contributions are summarized as follows:

- We develop a novel architecture with built-in detectors for open-vocabulary object detection through region-word alignment and built-in detector training.
- We propose an anchor proposal network that can adaptively provide simplified proposals based on images and queries, significantly accelerating the inference process.
- Extensive experiments conducted on two public benchmarks, COCO [16] and LVIS [9], demonstrate the effectiveness of our method.

2. Related Works

Open-Vocabulary Object Detection. OVR-CNN [32] proposed the concept of open vocabulary object detection (OVD). After observing that the Faster-RCNN [23] trained in a zero-shot manner tends to overfit to the base classes in OVD, OVR-CNN proposed to utilize a large-scale of external image-caption data to learn a rich vision-language space, and then apply it to the prediction of object categories during detection. Grad-OVD [6] proposed to generate pseudo bounding-box annotations with vision-language models for image-caption data, achieving region-level image-text matching. Detic [35] introduced ImageNet21K [4] source data without box annotations to train classifier branch and box prediction branch jointly, leveraging significantly richer image-text data. OvarNet [2] detected visual attributes with Faster-RCNN and CLIP-Attr, mining more fine-grained image-text information. These methods are designed based on frameworks with region localization ability, and then extend the close-set classification space to the open-vocabulary space by performing classification in vision-language space learned from a large-scale image-text pairs. Despite numerous advancements, the expansion from existing information in such methods often leads to inaccurate proposal-concept pairs, thereby constraining their ability to generalize to novel classes.

Transfering the Knowledge of Pre-Trained VLMs to Detection. Recent advancements in Vision-Language Models (VLMs), particularly CLIP [22], which is pre-trained on extensive image-text pair datasets, showcase remarkable zero-shot generalization aptitudes for vision-language

tasks. Building upon this foundation, modern OVD methods try to transfer the vision-language alignment ability of CLIP to detection models via knowledge distillation.

ViLD [8] trained RPN-based object detectors by distilling visual features from a pretrained CLIP model to the region-proposals. RegionCLIP [34] observed that CLIP has poor recognition ability on regions, and trained CLIP at region-word level with pseudo-labels generated from captioning. PromptDet [5] proposed learnable regional prompt to align regions-words embeddings extracted by CLIP. OADP [27] designed object-level distillation to obtain precise knowledge of objects as well as global and block distillation for more comprehensive knowledge transfer. BARON [29] grouped regions as a bag and aligned with the cropped region embeddings distilled from CLIP. CORA [30] proposed region prompting by designing a region classifier using RoIAlign and CLIP, followed by an anchor pre-matching object localization mechanism. VLDet [15] defined the extraction of region-word pairs from image-text pairs as an element matching problem of two sets which could be solved by binary matching.

OWL-ViT [18] designed DETR-based [1] framework, where object category names are utilized as queries for each image. It pretrained at image-text level, followed by fine-tuning the detector to generalize at region-text level. RO-ViT randomly cropped and resized regions of positional embeddings instead of using the whole image positional embeddings in the pretraining stage. OV-DETR [31] formulated the learning objective as a binary matching between input queries and the corresponding objects, which improve class-aware regression by conditionally validating queries against the text embedding. It uses conditional statements to determine whether a detection box matches the object. Specifically, for each detection box, the model calculates the matching score between it and the object, and uses a conditional statement to determine whether the score is above the threshold. If the score is above the threshold, the detection box will be considered as a matching object. Otherwise, consider it as the background.

OV-DETR is a model with high computational consumption. When dealing with samples with multiple class objectives, each class has to be processed separately, which further increases computational cost and is not conducive to model convergence. During training, the number of negative classes sampled in each iteration is limited due to the memory constraint, which hinders convergence. In the inference phase, repetitive per-class decoding is required, leading to low inference efficiency. Therefore, conditional matching has limitations and may not be suitable for large-scale models with a open vocabulary space. Different from OV-DETR, we designe an anchor proposal network to avoid redundant detection ports, thereby improving training and inference efficiency.

3. Method

3.1. Problem Definition

Given a corpus of image-text pairs that are associated with an open vocabulary space $\mathcal{C}_O$, we first learn the region-word alignment with a dual-encoder that is pre-aligned across different modalities (image-caption level). Then, we learn object localization and classification with a Transformer decoder under the supervision of categories and bounding box annotations of a base dataset where the vocabulary is limited to a base class space $\mathcal{C}_B$. During the inference phase, we aim to detect novel objects that belong to a novel class space $\mathcal{C}_N$. $\mathcal{C}_N$ is independent from $\mathcal{C}_B$, but may be explicitly or implicitly included in $\mathcal{C}_O$.

3.2. Region-word Alignment

Given an image-text pair $\langle I, T \rangle$, we extract the feature maps M of I with DINO-v2 [20] and obtain the region proposals using an off-the-shelf region proposal network (RPN) [23] simultaneously. Then, the proposed region embeddings are extracted with the help of RoIAlign [10]. We denote the region embeddings as $\boldsymbol{R} = \{r_1, r_2, ..., r_n\}$ where n is the number of regions. As for the corresponding caption T, we embed each noun of the caption with a text encoder and get a set of word embeddings $\boldsymbol{W} = \{w_1, w_2, ..., w_m\}$.

Given $\boldsymbol{R}$ and $\boldsymbol{W}$, we aim to find the best matching between r_i and w_j, that is to say, the best match region for a word, and vice versa. This can be formulated as a bipartite matching problem where the cost matrix $\boldsymbol{V}$ is defined as the inner products between the embeddings of regions and words, *i.e.*, $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{W}^{\mathsf{T}} \in \mathbb{R}^{n \times m}$. Therefore, the best matching problem is formulated by solving:

$$\begin{aligned} &\min_{\boldsymbol{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{i,j} S_{i,j} \\ &\text{s.t.} \quad S_{i,j} \in \{0, 1\} \end{aligned} \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{n \times m}$ is the matching relationship matrix between image regions and words. Concretely, $S_{i,j} = 1$ represents the i-th region r_i is matched with the j-th word w_j. The optimization problem 1 can be solved by the classical Hungarian matching algorithm [14].

After obtaining the matching relationship matrix $\boldsymbol{S}$, the fine-grained visual-language alignment space (region-word alignment) can be learned by classification loss. Furthermore, to alleviate the common problem of data imbalance in open vocabulary datasets, we leverage the focal loss [14], and the loss is calculated by:

$$\mathcal{L}_{focal} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} (1 - p_i)^{\gamma} \log(p_i), \tag{2}$$

where γ is the focusing parameter, p_i is the probability of

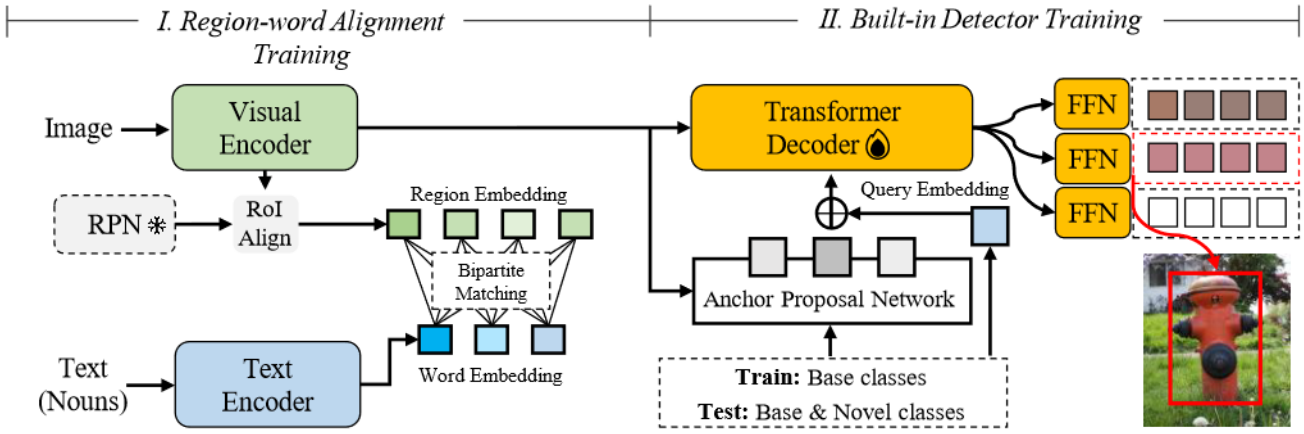

Figure 2. An overview of the proposed method. *I. Region-word Alignment.* We first extract the feature map of the image with a DINO-based ViT (pre-aligned in image-caption level). Then an off-the-shelf RPN is used to assist in getting region proposals ONLY at the region-word alignment training phase. The region representations are extracted by RoIAlign [10]. Meanwhile, we embed each noun of the caption with a text encoder, resulting in a set of word embeddings. We formulate the region-word alignment as a bipartite matching problem. *II. Built-in detector training.* The decoder is trained using annotated data, where the global features of the image are utilized as key and value, while the query embedding is augmented with positional embedding provided by the Anchor Proposal Network(Figure 3) to serve as a query.

true class:

$$p_i = \begin{cases} \sigma(r_i w_j/\tau) & \text{if } S_{i,j}=1 \\ 1-\sigma(r_i w_j/\tau) & \text{if } S_{i,j}=0 \end{cases}, \tag{3}$$

where σ is the sigmoid activation, τ is an adjustment parameter.

3.3. Built-in Detector Training

The alignment of region words described above serves as a fundamental requirement for our method. In this section, we introduce the construction of a built-in detector that efficiently locates objects by directly leveraging the aligned embeddings of image regions and words. Our detector adopts a DETR-style [1] Transformer structure, which has the capability to predict a fixed-size set of bounding boxes (x_i, y_i, h_i, w_i). Previous OVD methods, like OV-DETR [31], have also adopted the DETR-style structure. OV-DETR designed a binary matching loss to transform the category prediction in DETR into a binary classification problem, extending DETR to the task of open vocabulary object recognition.

However, this approach is hindered by high computational demands: the extensive set of bounding boxes leads to a significant amount of unnecessary calculations. Additionally, it struggles with processing multiple class queries simultaneously. To alleviate this dilemma, we propose an anchor proposal network, which can reduce computational complexity and support parallel training of samples that contain multiple classes of targets.

Anchor Proposal Network. Computational consumption is mainly caused by blind prediction of the target box during the training phase, which also leads to inefficiency in the inference phase. We propose an anchor proposal network to provide pre-selected proposals for the decoder. Specifically, during the training phase, given a pending image $\boldsymbol{x}$, we first use a visual encoder to extract the patch embedding of the image:

$$\boldsymbol{C} = f_\theta(\boldsymbol{x}) \in \mathbb{R}^{n\times d}. \tag{4}$$

Then we use a query encoder to extract the global feature of the query, where $\boldsymbol{q}$ represents the query, which can be text or a prompt image.

$$\boldsymbol{Q} = h_\theta(\boldsymbol{q}) \in \mathbb{R}^{1\times d}. \tag{5}$$

We obtain the matching possibility between each image patch and query through the interaction of query embedding and image patch embedding, and then use max pooling and normalization operations to obtain the positional filter of the anchor:

$$\boldsymbol{S} = \text{Normalize}\big(\text{MaxPool}(\boldsymbol{Q}C^{\mathsf{T}})\big) \in [0,1]^n \tag{6}$$

Finally, the anchor proposal is the non-zero vector in the Hadamard product of the learned positional embeddings $\boldsymbol{P}$ and the positional filter (We repeat d times for the feature dimension of $\boldsymbol{P}$ and denote it as $\hat{\boldsymbol{S}}$).

$$\boldsymbol{A} = \hat{\boldsymbol{S}}\boldsymbol{P}. \tag{7}$$

Anchor Matching Loss. After obtaining the anchor proposal, we add the query embedding with each anchor pro-

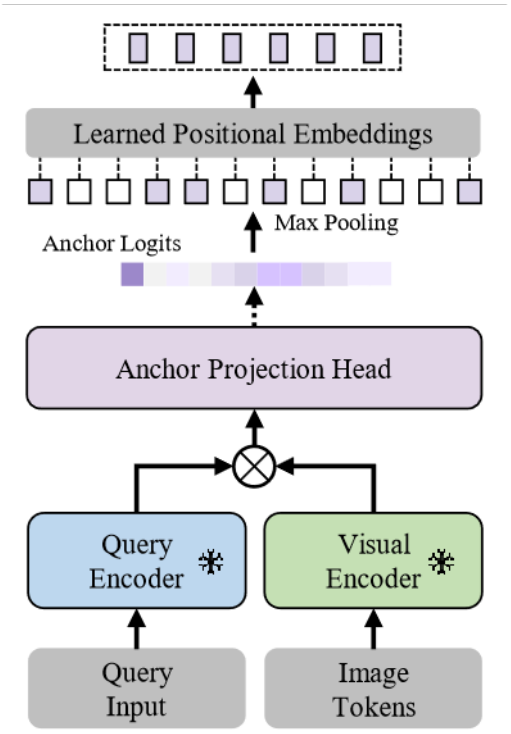

Figure 3. Anchor Proposal Network. We obtain the matching matrix through the interaction of query and visual embedding. Then max pooling and normalization operations are used to obtain the filter for the position of anchors.

posal as input to the Transformer decoder z_θ:

$$\hat{\boldsymbol{Q}} = \boldsymbol{Q} \oplus \boldsymbol{A}. \tag{8}$$

At the same time, the global features $\hat{\boldsymbol{C}}$ of the image are also input as keys and values to the decoder. Finally, a feed forward network (FFN) is used to predict the corresponding bounding box for each anchor proposal.

$$\hat{\boldsymbol{Y}} = \text{FFN}\left(z_\theta(\hat{\boldsymbol{C}}, \hat{\boldsymbol{Q}})\right), \tag{9}$$

where FFN is a 3-layer perceptron with ReLU activation function and a linear projection layer. $\boldsymbol{Y}$ is the bounding box predictions and each set of predictions corresponds to one anchor proposal. Following the prior methods, we add a special class label to represent the 'background', which means no object was detected. Now we have a set of predictions to match with a set of ground truth boxes. Following OV-DETR [31], we use a binary matching loss for annotation assignment:

$$\mathcal{L}_{\text{cost}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \mathcal{L}_{\text{match}}(\mathbf{p}, \hat{\mathbf{p}}) + \mathcal{L}_{\text{box}}(\mathbf{b}, \hat{\mathbf{b}}), \tag{10}$$

where $\mathcal{L}_{\text{match}}(\mathbf{p}, \hat{\mathbf{p}})$ is the binary classification loss between class prediction $\mathbf{p}$ and groud-truth $\hat{\mathbf{p}}$, which is implemented by the focal loss [14]. $\mathcal{L}_{\text{box}}(\mathbf{b}, \hat{\mathbf{b}})$ is the localization loss between localization prediction $\mathbf{b}$ and groud-truth $\hat{\mathbf{b}}$, which is implemented by the linear combination of L1 loss and a generalized IoU [33] loss for boxes. We optimize our model with the following objective:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}}. \tag{11}$$

Comparison with Anchor Pre-Matching [30]. Anchor pre-matching is also designed to provide accurate anchor boxes, which refers to the early matching of prediction boxes with ground truth boxes in the training of detection networks, specifying a category and regression target for each proposal, so that accurate category and position information can be learned more quickly in subsequent training. This process is usually achieved by calculating the IoU (Intersection over Union) of each proposal and all real target boxes. If the IoU is greater than a preset threshold, the proposal is matched with the corresponding real target box.

Despite the development, we argue it still has the following limitations: pre-matching adheres to strong assumptions that high embedding similarity means matching, which inevitably leads to incorrect matching and makes subsequent training uncorrectable, resulting in a preference for embedding similarity in the model and suboptimal performance.

In contrast, our method does not rely on strong assumptions but rather uses a specific network for position encoding inference. This preserves all regions while adaptively providing suitable proposals based on candidate images and queries, reducing redundancy and improving accuracy.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct evaluations on the two popular Open-Vocabulary object detection datasets, *i.e.*, LVIS [9] and COCO [16]. For the LVIS dataset which contains a large and diverse set of 1203 object categories, we follow previous work [8] to set the 337 rare categories as novel categories and leave the rest common and frequent categories into base categories. For the COCO dataset, we follow OVR-CNN [32] to divide the object categories into 48 base categories and 17 novel categories. Besides, the CC3M [26] dataset which contains 3 million image-text pairs is utilized in the pre-trained stage.

Evaluation Metrics. We evaluate the detection performance on both base and novel categories for completeness. For COCO, we follow OVR-CNN [32] to report the box AP at IoU threshold 0.5, noted as AP_{50}. For OV-LVIS, we report both the mask and box AP averaged on IoUs from 0.5 to 0.95, noted as mAP. The AP_{50} of novel categories ($\text{AP}_{50}^{\text{novel}}$) and mAP of rare categories (AP_r) are the main metrics that evaluate the open-vocabulary detection performance on OV-COCO and OV-LVIS, respectively. In addition, we also use AP_c, AP_f, and AP for common, frequent, and all categories in OV-LVIS.

Method	Backbone	Supervision	Built-in Detector	Generalized (17 + 48) Novel	Base	All
OVR-CNN [32]	ViT-B/32	Base	✗ (F-RCNN)	27.5	46.8	39.9
ViLD [8]	ViT-B/32	Base+Novel	✗ (F-RCNN)	27.6	59.5	51.3
Detic [35]	RN50	Base	✗ (F-RCNN)	27.8	47.1	45.0
OV-DETR [31]	ViT-B/32	Base+Novel	✓	29.4	61.0	52.7
Ro-ViT [13]	ViT-B/16	Base	✗ (OLN-RPN)	30.2	-	41.5
CFM-ViT [12]	ViT-B/16	Base	✗ (OLN-RPN)	30.8	-	42.4
RegionCLIP [34]	RN50	Base	✗ (F-RCNN)	31.4	57.1	50.4
MEDet [3]	RN50	Base	✗ (F-RCNN)	32.6	54.0	49.4
BARON [29]	RN50	Base	✗ (F-RCNN)	34.0	60.4	53.5
CORA [30]	RN50	Base	✓	35.1	35.5	35.4
OADP [27]	ViT-B/32	Base	✗ (F-RCNN)	35.6	55.8	50.5
BIND (Ours)	ViT-B/16	Base	✓	<u>36.3</u>	54.7	50.2
BIND (Ours*)	ViT-L/16	Base	✓	**41.5**	58.3	54.8

Table 1. Comparison with state-of-the-art methods on OV-COCO benchmark. F-RCNN reprsents Faster-RCNN [24]. OLN-RPN is [11]. **Best** and <u>second best</u> results are highlighted. Our method achieves state-of-the-art performance.

Method	Backbone	Supervision	Built-in Detector	AP_r	AP_c	AP_f	AP
ViLD [8]	ViT-B/32	Base+Novel	✗ (F-RCNN)	16.3	21.2	31.6	24.4
RegionCLIP [34]	RN50	Base	✗ (F-RCNN)	17.1	27.4	34.0	28.2
Detic [35]	RN50	Base	✗ (F-RCNN)	17.8	26.3	31.6	26.8
OV-DETR [31]	RN50	Base+Novel	✓	21.0	25.0	32.5	26.6
OADP [27]	ViT-B/32	Base	✗ (F-RCNN)	21.9	28.4	32.0	28.7
RegionCLIP [34]	RN50×4	Base	✗ (F-RCNN)	22.0	32.1	36.9	32.3
MEDet [13]	RN50	Base	✗ (F-RCNN)	22.4	-	-	34.4
BARON [29]	RN50	Base	✗ (F-RCNN)	23.2	29.3	32.5	29.5
Ro-ViT [13]	ViT-B/16	Base	✗ (OLN-RPN)	28.0	-	-	30.2
CORA [30]	RN50	Base	✓	28.1	-	-	-
CFM-ViT [12]	ViT-B/16	Base	✗ (OLN-RPN)	28.8	-	-	32.0
BIND (Ours)	ViT-B/16	Base	✓	<u>29.4</u>	30.6	33.5	31.4
BIND (Ours*)	ViT-L/16	Base	✓	**32.5**	33.4	35.3	33.2

Table 2. Comparison with state-of-the-art methods on OV-LVIS benchmark. F-RCNN reprsents Faster-RCNN [24]. OLN-RPN is [11]. **Best** and <u>second best</u> results are highlighted. Our method achieves state-of-the-art performance.

Method	Backbone	Novel AP_{50}	Secs/Img ↓
CORA [30]	RN50×4	41.7	0.50
BIND (Ours*)	ViT-L/16	41.5	0.33
BIND (Ours)	ViT-B/16	36.3	**0.21**

Table 3. Efficiency comparison with state-of-the-art methods on OV-COCO benchmark. Our method achieves 0.17 seconds acceleration in the inference with similar accuracy.

Method	Backbone	Novel AP_{50}	Secs/Img ↓
OWL-ViT† [19]	ViT-L/14	31.2	0.42
BIND (Ours*)	ViT-L/16	32.5	0.71
BIND (Ours)	ViT-B/16	29.4	**0.39**

Table 4. Efficiency comparison with state-of-the-art methods on OV-LVIS benchmark.† indicates the method is implemented with JAX framework.

Implementation Details. We use DINOv2 [20] ViT as the image encoder and a pre-trained language model CLIP [21] as our text encoder. We use a DETR-style Transformer decoder, which has 6 layers of width 256 with 8 attention heads. To prevent information leakage, RPN training is only conducted on the base class data. We only use RPN in train-

ing for stabilizing initial steps and **DO NOT** need it during the inference phase.

We first train the dual-encoder for cross-modal alignment using a CLIP [22]-style contrastive loss on pre training datasets for 12 epochs. Our cross modal alignment is a fine-tuning process of the DINOv2 [20] ViT visual encoder where only the last two layers are optimized and a pre-trained Transformer text encoder [21] where only an adaptation layer is optimized by the above image-caption level contrastive loss.

In region-word alignment training, the dual-encoder is optimized by Adam with a learning rate of 1e-4 and a weight decay of 1e-5. The parameters of the dual-encoder are fixed after region-word alignment pre-training, *i.e.*, only the decoder is optimized in the following object localizer training phase. We implement the proposed method with PyTorch and train the model for 50 epochs in the region-word alignment pre-training phase and another 50 epochs in the object localizer training phase. We set focusing parameter γ to 2, and adjustment parameter τ to 0.1 in the experiments. The weights of $\mathcal{L}_{\text{focal}}$, $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{GIoU}}$ is 1, 3 and 1 respectively.

4.2. Comparison with State-of-the-Art Methods

We compare with both common transfer-based methods and prior built-in detector methods. a) External Detector Methods: including Fast-RCNN (RegionCLIP [34], MEDet [3], BARON [29], OADP [27], *etc.*) based and OLN-RPN [11] (Ro-ViT [13], CFM-ViT [12]) based methods. b) Built-in Detector Methods: OV-DETR [31] and CORA [30].

COCO Benchmark. Table 1 shows the comparison results with prior methods on the open-vocabulary COCO benchmark. Our method achieves 36.3% AP_{50} on novel classes, which outperforms all prior methods in both external and built-in tracks. In addition, our method achieves comparable performance in both base and all class settings. It is worth noting that our method gains significant implementation in all metrics when using larger-scale backbones, indicating that our method has good scalability in model size.

LVIS Benchmark. Table 2 shows the comparison results with prior methods on the open-vocabulary LVIS benchmark. Our method achieves 29.4% AP_{50} on novel classes, which outperforms all prior methods in both ViT-based and ConvNet-based tracks. Furthermore, our method achieves comparable performance in both base and all class settings. We also tested the scalability of the model under this benchmark and larger models can bring performance gains.

4.3. Analysis

Efficiency. To demonstrate the efficiency of our method, we tested the inference speed of the models on two benchmark datasets, using the metric of seconds per image to detect an image. We conducted efficiency experiments on an A100, with both CORA [30] and our method implemented using Pytorch, while OWL-ViT [19] is implemented with JAX, a recognized framework with speed advantages. Table 3 and Table 4 shows the results of our comparison with those methods. On the COCO benchmark., when using the same larger model, we maintained close performance while significantly improving inference speed. When using smaller models, the inference speed can be further accelerated. On the LVIS benchmark, our inference speed is faster than OWL-ViT [19] when using smaller models. This is thanks to our designed anchor, which can provide reliable proposals and avoid computational redundancy.

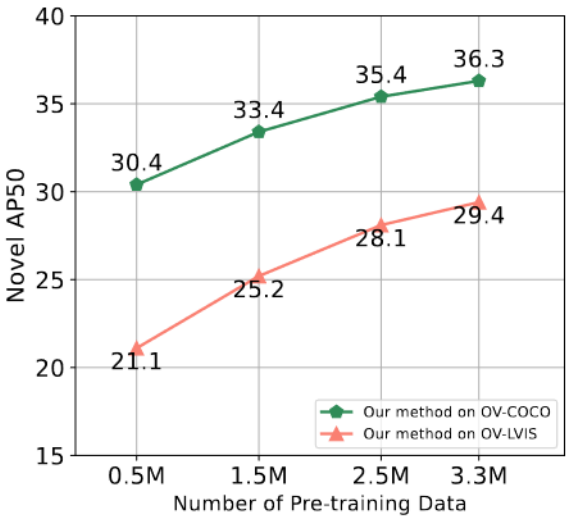

Figure 4. Scalability study with the different number of pre-training data pairs.

Scalability. To figure out how our method scales with the size of the pre-training dataset, we pre-train the model with 0.5M, 1.5M, 2.5M, and 3.3M data pairs successively. We report the AP_{50} of Novel classes on OV-COCO and mAP of rare categories (AP_r) in Figure 4, As the amount of data increases, the performance of the model continues to increase, and the improvement is more significant at the beginning, which demonstrates its good scalability. It may be attributed to the design of region-word alignment.

Visualization. To evaluate the quality of the proposed object positions given by our anchor proposal network, we randomly selected four images for visualization. Each image is resized to 224×224, and the patch size is 16×16, resulting in 14×14 patches per image. The proposed patches are denoted with a translucent mask. From Figure 5, it can be seen that our anchor proposal network can provide high-quality object positions.

4.4. Ablation Study

We conduct ablation studies on the OV-COCO dataset to reveal the effectiveness of each component in our proposed framework.

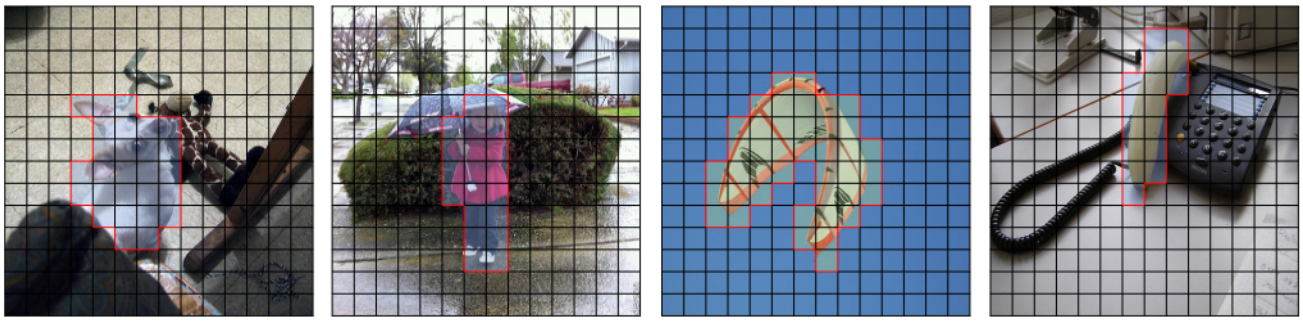

Figure 5. The visualization of regions generated by our anchor proposal network. The frames of the positional proposal are highlighted in red. The image is resized to 224×224, the patch size is 16×16.

Pre-train Model	Backbone	Novel AP_{50}	ALL AP_{50}
CLIP	RN50	33.8	45.4
CLIP	ViT-B/16	34.2	46.6
DINOv2	ViT-B/16	36.3	50.2

Table 5. Ablation studies of DINOv2 [20] ViT visual encoder. DINOv2 pre-train model achieves better performance.

RPN	Backbone	Novel AP_{50}	Secs/Img ↓
Prediction Slot	ViT-B/16	32.4	0.63
BIND (Ours)	ViT-B/16	**36.3**	**0.21**

Table 6. Ablation studies of anchor proposal network. *Prediction Slot* is an anchor proposal method in DETR [31]-style.

Pre-train Model	R-W	Novel AP_{50}	ALL AP_{50}
CLIP (R50)	✓	33.8	45.4
	✗	24.5	27.6
CLIP (ViT-B/16)	✓	34.2	46.6
	✗	25.1	28.5
DINOv2 (ViT-B/16)	✓	36.3	50.2
	✗	28.4	31.7

Table 7. Ablation studies of DINOv2 [20] ViT visual encoder. "R-W" represent region-word alignment. Removing region-word alignment at different Pre-train Models result in a significant decrease in performance.

Pre-train Model. In order to explore the impact of different pre-train models on our method, we employ CLIP in our framework, which includes two types of backbones: ViT-based and ConvNet-based. As shown in Table 5, We achieve better performance with DINOv2 [20] ViT visual encoder. This confirms our motivation that self-supervised trained DINOv2 may have mined intrinsic association within the image, which is more suitable for region pre-training representations to align with word than VLM.

Region-word Alignment. To verify the importance of region-word alignment, we use different pre-train models while removing the training for Region word alignment. As shown in Table 7, whether it is CLIP or our method, abandoning region-word alignment will result in significant performance degradation. It is worth noting that our method has the smallest reduction. This ablation study demonstrates that region word alignment is a crucial part for the open-vocabulary object detection task.

Anchor Proposal Network. In order to evaluate the effectiveness of the anchor proposal network, we used a DETR [1] style during the inference phase, providing prompt information for the object location by setting many prediction slots (set to 100 following with DETR [1]). In contrast, our anchor proposal network serves as a filter for prediction slots and most meaningless proposals will be discarded to reduce computational redundancy. As shown in Table 6, our method not only achieves good accuracy but also has significant advantages in speed.

5. Conclusion

In this paper, we present a novel architecture for open-vocabulary object detection that features a built-in detector, obviating the need for module replacement or knowledge transfer. Our two-stage training framework, consisting of an image-text dual-encoder and a DETR-style decoder, demonstrates an Encoder-Decoder structure. The former learns region-word alignment from a corpus of image-text pairs, while the latter performs detection on annotated object detection datasets. In contrast to traditional manually designed non-adaptive anchors, our anchor proposal network generates high-likelihood anchor proposals based on candidates adaptively, significantly improving detection efficiency. Empirical evaluations on the COCO and LVIS benchmarks attest to our method's status as a state-of-the-art approach to open-vocabulary object detection.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. *End-to-End Object Detection with Transformers*, pages 213–229. 2020. 1, 2, 3, 4, 8

[2] K. Chen, X. Jiang, Y. Hu, X. Tang, Y. Gao, J. Chen, and W. Xie. Ovarnet: Towards open-vocabulary object attribute recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23518–23527, 2023. 2

[3] Peixian Chen, Kekai Sheng, Mengdan Zhang, Yunhang Shen, Ke Li, and Chunhua Shen. Open vocabulary object detection with proposal mining and prediction equalization. *CoRR*, abs/2206.11134, 2022. 6, 7

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[5] 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. 3

[6] 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

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 3, 5, 6

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3, 4

[11] 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*, PP:1–1, 2021. 6, 7

[12] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15602–15612, 2023. 6, 7

[13] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154, 2023. 6, 7

[14] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. 3, 5

[15] 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. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 5

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[18] Austin Stone Maxim Neumann Dirk Weissenborn Alexey Dosovitskiy Aravindh Mahendran Anurag Arnab Mostafa Dehghani Zhuoran Shen Xiao Wang Xiaohua Zhai Thomas Kipf Neil Houlsby Matthias Minderer, Alexey Gritsenko. Simple open-vocabulary object detection with vision transformers. 2022. 2, 3

[19] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, 2022. 6, 7

[20] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 3, 6, 7, 8

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 6, 7

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 7

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2, 3

[24] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 6

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[26] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5

[27] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11186–11196, 2023. 2, 3, 6, 7

[28] Jianzong Wu, Xiangtai Li, Shilin Xu Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, et al. Towards open vocabulary learning: A survey. *arXiv preprint arXiv:2306.15880*, 2023. 1, 2

[29] 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, 3, 6, 7

[30] 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 (CVPR)*, pages 7031–7040, 2023. 2, 3, 5, 6, 7

[31] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *ArXiv*, abs/2203.11876, 2022. 3, 4, 5, 6, 7, 8

[32] 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 (CVPR)*, pages 14393–14402, 2021. 1, 2, 5, 6

[33] Hongyu Zhai, Jian Cheng, and Mengyong Wang. Rethink the iou-based loss functions for bounding box regression. *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 9:1522–1528, 2020. 5

[34] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chengkun Li, Noel C. F. Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16772–16782, 2022. 2, 3, 6, 7

[35] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 6