CORA: Adapting CLIP for Open-Vocabulary Detection with Region Prompting and Anchor Pre-Matching

Xiaoshi Wu[1], Feng Zhu[2], Rui Zhao[2,3], Hongsheng Li[1,4]
[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]SenseTime Research [3]Qing Yuan Research Institute, Shanghai Jiao Tong University
[4]Centre for Perceptual and Interactive Intelligence (CPII)
{wuxiaoshi@link, hsli@ee}.cuhk.edu.hk, {zhufeng, zhaorui}@sensetime.com

Abstract

*Open-vocabulary detection (OVD) is an object detection task aiming at detecting objects from novel categories beyond the base categories on which the detector is trained. Recent OVD methods rely on large-scale visual-language pre-trained models, such as CLIP, for recognizing novel objects. We identify the two core obstacles that need to be tackled when incorporating these models into detector training: (1) the distribution mismatch that happens when applying a VL-model trained on whole images to region recognition tasks; (2) the difficulty of localizing objects of unseen classes. To overcome these obstacles, we propose CORA, a DETR-style framework that adapts **C**LIP for **O**pen-vocabulary detection by **R**egion prompting and **A**nchor pre-matching. Region prompting mitigates the whole-to-region distribution gap by prompting the region features of the CLIP-based region classifier. Anchor pre-matching helps learning generalizable object localization by a class-aware matching mechanism. We evaluate CORA on the COCO OVD benchmark, where we achieve 41.7 AP50 on novel classes, which outperforms the previous SOTA by 2.4 AP50 even without resorting to extra training data. When extra training data is available, we train CORA$^+$ on both ground-truth base-category annotations and additional pseudo bounding box labels computed by CORA. CORA$^+$ achieves 43.1 AP50 on the COCO OVD benchmark and 28.1 box APr on the LVIS OVD benchmark. The code is available at https://github.com/tgxs002/CORA.*

1. Introduction

Object detection is a fundamental vision problem that involves localizing and classifying objects from images. Classical object detection requires detecting objects from a closed set of categories. Extra annotations and training are required if objects of unseen categories need to be detected. It has attracted much attention on detecting novel categories without tedious annotations, or even detect object from new category, which is currently referred as open-vocabulary detection (OVD) [36].

Recent advances on large-scale vision-language pre-trained models, such as CLIP [30], enable new solutions for tackling OVD. CLIP learns a joint embedding space of images and text from a large-scale image-text dataset, which shows remarkable capability on visual recognition tasks. The general idea of applying CLIP for OVD is to treat it as an open-vocabulary classifier. However, there are two obstacles hindering the effective use of CLIP on tackling OVD.

How to adapt CLIP for region-level tasks? One trivial solution is to crop regions and treat them as separate images, which has been adopted by multiple recent works [7, 14, 31, 35]. But the distribution gap between region crops and full images leads to inferior classification accuracy. MEDet [7] mitigates this issue by augmenting the text feature with image features. However, it requires extra image-text pairs to prevent overfitting to so-called “base” classes that are seen during training. RegionCLIP [40] directly acquires regional features by RoIAlign [17], which is more efficient but cannot generalize well to novel classes without finetuning. The finetuning is costly when adopting a larger CLIP model.

How to learn generalizable object proposals? ViLD [14], OV-DETR [35], Object-Centric-OVD [31], RegionCLIP [40] need RPN or class-agnostic object detectors [29] to mine potential novel class objects. However, these RPNs are strongly biased towards the base classes on which they are trained, while perform poorly on the novel classes. MEDet [7] and VL-PLM [39] identify this problem and adopt several handcrafted policies to rule out or merge low-quality boxes, but the performance is still bounded by the frozen RPN. OV-DETR [35] learns generalizable object localization by conditioning box regression on class name embeddings, but at the cost of efficiency issue induced by repetitive per-class inference.

In this work, we propose a new framework based on DEtection TRansformers (DETR) [6] that incorporates CLIP into detector training to achieve open-vocabulary detection without additional image-text data. Specifically, we use a DETR-style object localizer for class-aware object localization, and the predicted boxes are encoded by pooling the intermediate feature map of the CLIP image encoder, which are classified by the CLIP text encoder with class names. However, there is a distribution gap between whole-image features from CLIP's original visual encoder and the newly pooled region features, leading to an inferior classification accuracy. Thus, we propose Region Prompting to adapt the CLIP image encoder, which boosts the classification performance, and also demonstrates better generalization capability than existing methods. We adopt DAB-DETR [26] as the localizer, in which object queries are associated with dynamic anchor boxes. By pre-matching the dynamic anchor boxes with the input categories before box regression (Anchor Pre-Matching), class-aware regression can be achieved without the cost of repetitive per-class inference.

We validate our method on COCO [24] and LVIS v1.0 [16] OVD benchmarks. On the COCO OVD benchmark, our method improves AP50 of novel categories over the previous best method [40] by 2.4 AP50 without training on extra data, and achieves consistent gain on CLIP models of different scales. When compared under a fairer setting with extra training data, our method significantly outperforms the existing methods by 3.8 AP50 on novel categories and achieves comparable performance on base categories. On the LVIS OVD benchmark, our method achieves 22.2/28.1 APr with/w.o. extra data, which significantly outperforms existing methods that are also trained with/w.o. extra data. By applying region prompting on the base classes of COCO, the classification performance on the novel classes is boosted from 63.9% to 74.1%, whereas other prompting or adaptation methods easily bias towards the base classes.

The contributions of this work are summarized as follows: (1) Our proposed region prompting effectively mitigates the gap between whole image features and region features, and generalize well in the open-vocabulary setting. (2) Anchor Pre-Matching enables DETR for generalizable object localization efficiently. (3) We achieve state-of-the-art performance on COCO and LVIS OVD benchmarks.

2. Related Works

Open-Vocabulary Object Detection OVR-CNN [36] firstly put forth this new formulation of detection, and proposes its baseline solution by aligning region features with nouns in captions that are paired with the image. Mingfeng [11] et al. mine pseudo labels by utilizing the localization ability of pre-trained vision-language models. PromptDet [10] addresses the gap between image and region classification by adding learnable prompts when encoding the class names, namely regional prompt learning (RPL), which is expected to generalize from base to novel categories. OV-DETR [35] is the first DETR-style open-vocabulary detector, which proposes conditional matching to solve the missing novel class problem in assignment, but at the cost of inefficient inference. RegionCLIP [40] propose a second-stage pre-training mechanism to adapt the CLIP model to encode region features, and demonstrates its capability on OVD and zero-shot transfer setting. GLIP [21] jointly learns object localization and VL alignment. Matthias et al. [15] proposes to finetune a VL aligned model for detection, while we fix the pre-trained VL model for better generalization towards novel categories.

Detection Transformers DETR [6] is an object detection architecture based on transformers that formulates object detection as a set-to-set matching problem, which greatly simplifies the pipeline. Several works address the slow convergence problem of DETR by architectural improvement [1, 13, 26, 38] or special training strategies [8, 18]. Zhu et al. [1] proposes multi-scale deformable attention module to efficiently aggregate information from multi-scale feature maps. Gao et al. [13] proposes to modulate the cross-attention in the transformer decoder by anchor box coordinates to accelerate the detector convergence. DAB-DETR [26] formulates the queries in DETR architecture as anchor boxes, which accelerates detector training. Chen et al. [38] proposes Group DETR, which adds auxiliary object queries during training to take advantage of one-to-many matching for faster convergence.

Prompt Tuning Prompting is originated from NLP, and it refers to prepending task instructions before the input sequence to give the language model the hint about the task [5]. Later works [22, 27] explores tuning continuous prompt vectors when few-shot data is available. VPT [19], Visual Prompting [2, 3] explore prompting in the pixel space. [20] and [25] prompts pre-trained model for video recognition tasks. [34] proposes class-aware visual prompt tuning to generalize the learned prompts to unseen categories. Recent works demonstrate that prompt tuning is an effective and parameter-efficient way to adapt large-scale pre-trained models to downstream tasks.

3. Method

Open-vocabulary detection (OVD) is an object detection task aiming at detecting objects from novel categories beyond the base categories on which the detector is trained. Formally, the detector is trained on a detection dataset with base-category C^B box annotations, and tested on a new input image $I \in \mathbb{R}^{H \times W \times 3}$ to detect objects belonging to a novel category set C^N, where $C^B \cap C^N = \emptyset$. In this section, we introduce CORA, a framework that adapts CLIP for the OVD task by Region prompting and Anchor pre-matching. For fairer comparisons with existing methods, we also experiment with a broader setting where extra data is available, which is referred to as CORA$^+$, and will be introduced along with the experiments.

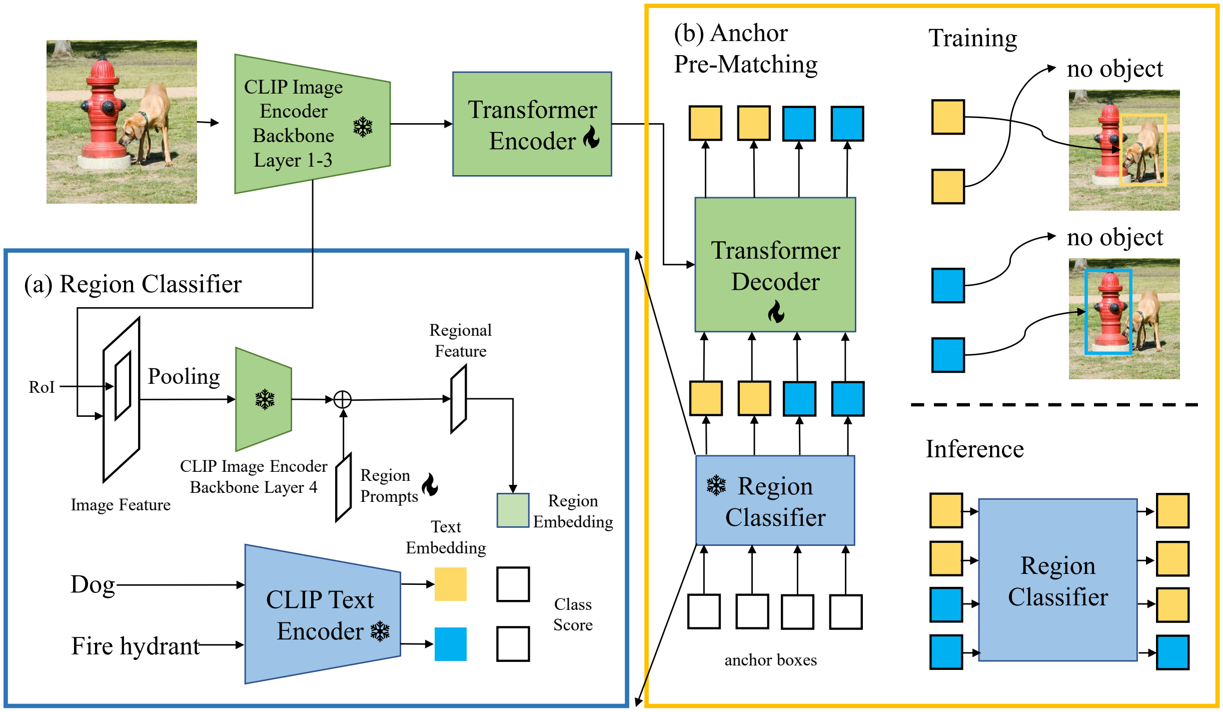


Figure 1. Overview of our method. The image is encoded into a feature map by the CLIP image encoder for both localization and classification. The regional feature is extracted by pooling the feature map, and then prompted before classified by the CLIP class name embeddings. The anchor boxes are pre-matched and conditioned on a class before decoding. During training, per-class post-matching is conducted. During inference, the box predictions are classified by the region classifier.

3.1. Overview

The overall framework of CORA is illustrated in Fig. 1. Given an image as input, we acquire the spatial feature map using the ResNet backbone from the pre-trained CLIP image encoder, which is shared by both region classification and object localization branches. Unlike conventional detectors, localization and classification are decoupled and sequentially conducted in our framework to better fit the characteristic of the OVD problem. We train a DETR-style object localizer that refines a set of object queries together with their associated anchor boxes to localize the objects, which are then classified by a region classifier adapted from CLIP.

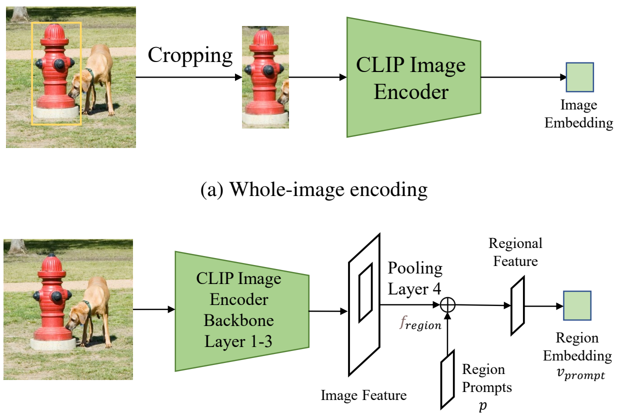
Region Classification. Given a region to be classified (anchor box or box prediction), we adopt RoIAlign to obtain the region feature, followed by the attention pooling module of CLIP to generate region embeddings, which can be classified by class embeddings obtained from the CLIP text encoder, as done in CLIP. We name this module as CLIP-based region classifier (Fig. 1-(a)).

Object Localization. The visual feature map is firstly refined by the DETR-like encoder, and then fed into the DETR-like decoder. The queries of anchor boxes are firstly classified by the CLIP-based region classifier, which are then conditioned on their predicted labels before iteratively refined by the DETR-like decoder for better localization. The decoder also estimates the matchability of the query with the previously predicted label. During training, the predicted boxes are one-to-one matched with ground truth boxes that has the same labels (Fig. 1-(b)), and trained as in DETR. In the inference stage, the class labels of the boxes are adjusted by the CLIP-based region classifier.

As mentioned in Sec. 1, there are two obstacles to be addressed: (1) object detection conducts recognition on image regions, while the CLIP model is trained on whole-image input, leading to a distribution gap that hinders classification performance. (2) the detector needs to learn object localization for novel classes, while we only have annotations on a limited number of base classes. To solve the first obstacle, we propose region prompting to modulate the region features for better generalizable region embeddings, which will be introduced in Sec. 3.2. To solve the second obstacle, we put forward anchor pre-matching to encourage class-aware object localization that can generalize to novel classes during inference, which will be introduced in Sec. 3.3.

3.2. Region Prompting

OVD requires the detector to classify image regions into a given category list. In this section, we will elaborate how



(a) Whole-image encoding

(b) Regional encoding

Figure 2. Comparison between the CLIP-based region classifier and the vanilla pipeline. We pool regional feature from the feature map, rather than cropping the region patch and classify them as a separate image.

a pre-trained CLIP model is adapted to formulate the CLIP-based region classifier. Given the CLIP model, region classification can be realized by comparing the similarity between the regional embedding from the CLIP image encoder and the class name embedding from the CLIP text encoder.

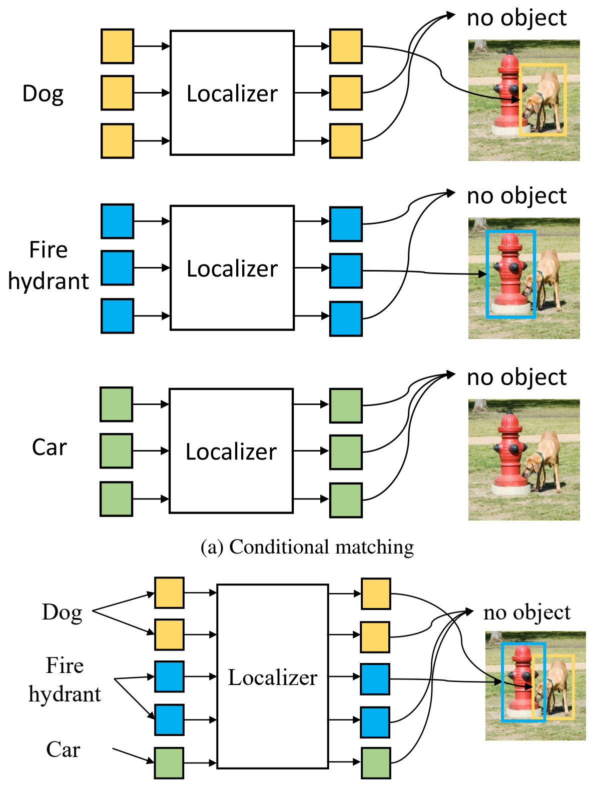
Region Prompting. As illustrated in Fig. 2, given an image and a set of region of interest (RoI), we firstly encode the whole image into a feature map by the CLIP encoder's first 3 blocks, which are then pooled by RoIAlign [32] either according to anchor boxes or predicted boxes into region features, before encoded by the last block of the CLIP image encoder backbone. There exists a distribution gap between the CLIP image encoder's whole-image feature map and the pooled regional features. We propose region prompting to fix the misalignment by augmenting the region feature with learnable prompts $p \in \mathbb{R}^{S \times S \times C}$, where S is the spatial size of the regional feature, and C is the dimension of the regional feature. Specifically, given the input regional feature f_{region}, the region prompting is conducted as

$$v_{\text{prompt}} = P(f_{\text{region}} \oplus p), \tag{1}$$

where $\oplus$ denotes element-wise addition, P is the attention pooling module of the CLIP visual encoder.

Optimizing Region Prompts. We train the region prompts on a detection dataset with base-class annotations. The class name embeddings are pre-computed by the CLIP text encoder, which are used as classifier weights later. We train the prompts by a standard cross-entropy loss to classify the ground truth boxes with their pooled regional features f_{region}. When optimizing the region prompts, we keep other model weights frozen, and only make the region prompts to be learned.

Comparison with existing methods. The previous common practice for region classification with CLIP is to crop the RoIs, and encode them as separate images, before comparing with text embeddings. This pipeline is not efficient when encoding regions with overlaps, since the overlapping regions are encoded more than once in different region crops. Its accuracy also suffers from the missing context information. In contrast, our regional prompting is more efficient and preserves richer context.



(a) Conditional matching

(b) Anchor pre-matching

Figure 3. Comparison between anchor pre-matching and conditional matching. Anchor pre-matching decodes a constant number of object queries, and assigns different numbers of object queries based on the image content, avoiding repetitive decoding as in conditional matching.

The region prompts contain less than 1M parameters, which is in line with recent advances of the prompt tuning and adapter literatures. Region prompting generalizes well to unseen novel classes. We attribute the generalization capability to the fact that region prompting directly **fix** the distribution mismatch right after where it occurs (after region pooling), whereas the existing methods tweak irrelevant parameters to **compensate** for the distribution mismatch.

3.3. Anchor Pre-Matching

Region prompting helps solve the region classification problem. Object localization is the other critical subtask of object detection. Considering the inferior performance of pre-trained RPN on novel classes, we introduce a class-aware query-based object localizer, which demonstrates better generalization capability on unseen classes. As shown in Fig. 1, Given the visual feature map from the frozen CLIP image encoder, the object queries are pre-matched to the class name embeddings by the CLIP text encoder.

Anchor Pre-Matching. The object localizer is implemented by a DETR-style transformer encoder-decoder structure, where the encoder refines the feature map, and the decoder decodes a set of object queries into box predictions. We adopt DAB-DETR [26], where each object query is associated with an anchor box. Each ground truth box is pre-matched to a set of queries with the same label. The label $\hat{c}_i$ of an object query is assigned by classifying the associated anchor box b_i

$$\hat{c}_i = \arg\max_{c \in C^B} \text{cosine}(v_i, l_c), \tag{2}$$

where v_i is the region feature of anchor box b_i, l_c is the class name embedding of class c, and cosine denotes the cosine similarity. After pre-matching, each object query conditions on the predicted class embedding to allow class-aware box regression. The conditioned object query is given by

$$q_i = \text{MLP}(l_{\mathrm{c_i}}). \tag{3}$$

The DETR-like decoder iteratively refines each object query with its associated anchor box (q_i, b_i) into $\hat{y}_i = (\hat{p}_i, \hat{b}_i)$, where $\hat{b}_i$ is the refined box coordinates and $\hat{p}_i$ is the matching probability to the query's pre-matched class $\hat{c}_i$.

Given the model predictions, the assignment between ground truth boxes and model predictions is conducted by performing bipartite matching for each class separately. We only allow each ground truth box to be assigned to the prediction with the same pre-matched label, in order to enforce the decoder to be aware of the conditioned text embedding.

Specifically, for class c, given the N^c box predictions $\hat{y}^c = \{\hat{y}_i \mid \hat{c}_i = c\}$ that are pre-matched to class c, and the set of ground truth boxes y^c in class c, we optimize a permutation of N^c elements $\sigma \in \mathfrak{S}_{\mathrm{N^c}}$ that minimizes the following cost

$$\hat{\sigma}_c = \arg\max_{\sigma \in \mathfrak{S}_{\mathrm{N^c}}} \sum_{i}^{N^c} \mathcal{L}_{\text{cost}}(y_i^c, \hat{y}^c_{\sigma(\mathrm{i})}), \tag{4}$$

where the matching cost is defined as

$$\mathcal{L}_{\text{cost}}(y, \hat{y}) = \mathcal{L}_{\text{match}}(p, \hat{p}) + \mathcal{L}_{\text{box}}(b, \hat{b}). \tag{5}$$

$\mathcal{L}_{\text{match}}(p, \hat{p})$ is a binary classification loss, and $\mathcal{L}_{\text{box}}(b, \hat{b})$ characterizes the localization error of $\hat{b}$ w.r.t b. In our case, we implement $\mathcal{L}_{\text{match}}$ by the focal loss [23]. $\mathcal{L}_{\text{box}}$ is implemented by a weighted sum of L_1 loss and GIoU [37] loss following prior works.

The model is optimized by the following loss

$$\begin{aligned}\mathcal{L} &= \sum_{c \in C^B} \mathcal{L}_{\text{match}}(p^c, \hat{p}^c_{\hat{\sigma}_c}) + \mathcal{L}_{\text{box}}(b^c, \hat{b}^c_{\hat{\sigma}_c}) \\ &= \lambda_{\text{focal}} \mathcal{L}_{\text{focal}} + \lambda_{\mathrm{L_1}} \mathcal{L}_{\mathrm{L_1}} + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}.\end{aligned} \tag{6}$$

During inference, we adopt the region classifier introduced in Sec. 3.2 to classify the predicted boxes $\{\hat{b}_i\}$ for better classification accuracy. The class score is multiplied by the pre-matching score to account for the box quality

$$\mathrm{P}(\hat{b}_i \in c) = \hat{p}_i \text{cosine}(\hat{v}_i, l_c). \tag{7}$$

Comparison with Conditional Matching [35]. Conditional Matching in OV-DETR [35] also proposes to condition the queries on the text embedding for class-aware regression. But it suffers from repetitive per-class inference. Specifically, as shown in Fig. 3, each class in C^N needs to be separately localized with the same group of query, which means both computation and memory consumption scales linearly with the vocabulary size. During training, the number of negative classes sampled in each iteration is limited due to the memory constraint, which hinders convergence. During inference, repetitive per-class decoding is required and results in low inference efficiency, especially when the vocabulary size is large.

Contrary to conditional matching, our anchor pre-matching mechanism assigns anchor boxes for different classes adaptively according to the image content, which ensures a constant number of query decoupled from the category size. By anchor pre-matching, all the classes can be decoded together in one pass, eliminating the need for repetitive per-class decoding.

To improve the generalization capability and training convergence of open-vocabulary detectors equipped with Anchor Pre-Matching, we also introduce two effective training techniques, namely, “Drop Class” and “CLIP-Aligned Labeling”.

Class Dropout. The generalization capability of the model can be further boosted by randomly dropping categories during training. Since our goal is to train a detector that can detect objects from a user specified category list, training on a fixed list of categories leads to bias. We mitigate this bias by randomly dropping out the base categories during training. For efficiency reason, we implement this idea by splitting the base classes into two complementary groups and train on both of them, instead of training on one group while dropping the other group. Specifically, in each training iteration, we split C^B and the ground truth boxes at a probability of p, and train the detector on the same image with two complementary sets of categories. It enforces the model to condition its prediction on the query categories. Since a ground truth box in one group does not appear in the other group, the model needs to be aware of the categories to treat differently for the two sets of annotations.

CLIP-Aligned Labeling. Directly training the localizer on the original COCO dataset suffers from convergence issue.

By anchor pre-matching mechanism, a ground truth box incorporates training only when at least one query with the same pre-matched label exists. Otherwise, it is ignored, which hinders convergence. This issue can be partially attributed to the inaccurate anchor box. However, even if a ground truth box has an accurate anchor box, it may still be ignored due to the limited recognition accuracy limitation of the region classifier, or in other words, the ground truth box label is not aligned with the CLIP region classifier used for pre-matching. Thus, we relabel the boxes in the training dataset with the region classifier, which we refer to as CLIP-Aligned labeling. With this technique, more ground truth boxes can be matched.

4. Experiments

In this section, we comprehensively evaluate our CORA on the open-vocabulary detection task. Datasets and evaluation protocols are introduced in Sec. 4.1, and implementation details of our method are provided in Sec. 4.2. We compare with state-of-the-art methods in Sec. 4.3, demonstrating advantages of our CORA, and then validate the effectiveness of the proposed Region Prompting and Anchor Pre-Matching in Sec. 4.4 and Sec. 4.5, respectively.

4.1. Dataset & Training & Evaluation

Following the convention of the COCO OVD benchmark proposed in [4], the 80 classes in the COCO dataset [24] classes are divided into 48 base classes and 17 novel classes. The model is trained on the 48 base classes, which contains 107,761 images and 665,387 instances. The model is then evaluated on the validation set of novel classes, which contains 4,836 images and 33,152 instances from both the 48 base classes and the 17 novel classes. We also conduct experiments on the LVIS v1.0 [16] dataset. On LVIS dataset, the model is trained on 461 common classes and 405 frequent classes, which contains 100,170 images and 1,264,883 instances. After training, the model is evaluated on the LVIS validation set, which contains 19,809 images and 244,707 instances.

To make fairer comparisons with other methods, we propose CORA$^+$, which utilizes extra dataset or target novel class names. CORA$^+$ is a detector trained with both ground-truth base-category annotations and additional pseudo bounding box labels. When target class names are provided, we generate pseudo boxes of novel classes on the base training dataset using CORA, and when extra image-text dataset (text could be captions or class names) is available, pseudo boxes of all objects mentioned by the corresponding text are generated. We use standard detector architecture and training target to train CORA$^+$. SAM-DETR [38] is used for the experiment on COCO and CenterNet2 [43] is used for the experiment on LVIS.

In the OVD task, we evaluate our model under the "generalized" setting, in which the model needs to predict objects from both base and novel classes, and then evaluated on novel objects. On the COCO benchmark, we take AP50 as our evaluation metric, which counts the average precision at an intersection over union (IoU) of 50% for each class, and then averages among all the classes. For the LVIS OVD benchmark, we evaluate on the full validation dataset, and report the mean AP of boxes from the novel classes to compare with prior works [40]. For the region classification task, we let the model classify the ground truth boxes in the COCO dataset, the performance is evaluated in terms of mAP.

4.2. Implementation Details

Model Specifications. We use DAB-DETR [26] as the object localizer. Specifically, localizer is configured to have 1,000 object queries, 3 encoder layers and 6 decoder layers. We use a multi-layer perceptron (MLP) with 128 hidden neurons to transform class name embeddings into an object query. Following CLIP [30], each class embedding is computed as the average text embedding of the class name over 80 context prompts. In class dropout, each class is randomly assigned to one of the two groups with equal probabilities. When adopting region prompting on LVIS, classes in the common and frequent group are sampled with equal weights, meaning that the objects in the less frequent classes are over sampled. When training our method on LVIS, we sample 100 categories (including the ground truth categories) in each iteration. Since the number of classes in the LVIS dataset is much larger than that of COCO, we relax the matching constraint in anchor pre-matching, such that ground truth boxes can be post-matched to the anchor boxes with a similar label. Specifically, classes with a cosine similarity greater than 0.7 are considered similar.

Training & Hyperparameters. We train the region prompts for 5 epochs with a base learning rate of 10^{-4}, which decays after the 4^{th} epoch by a factor of 0.1. The localizer is trained for 35 epochs with a learning rate of 10^{-4} without learning rate decay. Both the region prompts and the localizer are trained with batch size 32 by the AdamW optimizer [28] with 10^{-4} weight decay. We apply gradient clipping with a maximal norm of 0.1. To stabilize training, we evaluate on the exponential moving average (EMA) of the model after training. The class dropout probability p is set as 0.2. λ_{focal}, λ_{L_1} and λ_{GIoU} are set as 2.0, 5.0, 2.0, respectively. For the experiment on LVIS, we use repeat factor sampling [16] with default hyperparameters to balance the training samples. We use non-maximum suppression (NMS) with an IoU threshold 0.5 during inference.

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

Tab. 1 summarizes our main results. Since the pre-trained model is crucial to the open-vocabulary capability of the detector, we compare our method with baseline methods that are trained with the same CLIP model. When compared with methods trained on CLIP RN50, CORA outperforms VL-PLM by 0.7 AP50 on novel classes. With a larger pre-trained model, our method improves the previous state-

Method	Detector Training			Generalized (17 + 48)		
	Extra Dataset	Pre-train Model	Require Novel Class	Novel	Base	All
OVR-CNN [36]	COCO Captions [9]	-	✗	22.8	46.0	39.9
Detic [42]	COCO Captions [9]	CLIP (text encoder)	✗	27.8	47.1	45.0
RegionCLIP [40]	CC3M [33]	CLIP (RN50)	✗	31.4	57.1	50.4
VL-PLM [39]	-	CLIP (RN50)	✓	34.4	60.2	53.5
CORA (Ours)	-	CLIP (RN50)	✗	**35.1**	35.5	35.4
ViLD [14]	-	CLIP (ViT-B/32)	✓	27.6	59.9	51.3
OV-DETR [35]	-	CLIP (ViT-B/32)	✓	29.4	61.0	52.7
MEDet [7]	COCO Captions [9]	CLIP (ViT-B/32)	✗	32.6	54.0	49.4
RegionCLIP [40]	CC3M [33]	CLIP (RN50x4)	✗	39.3	61.6	55.7
CORA (Ours)	-	CLIP (RN50x4)	✗	41.7	44.5	43.8
CORA$^+$ (Ours)	COCO Captions [9]	CLIP (RN50x4)	✗	**43.1**	60.9	56.2

Table 1. Main results on the COCO OVD benchmark. We report AP50 as the evaluation metric. The baseline methods are grouped by their pre-trained model. We also list the extra dataset requirement of each method, and whether they require the novel class to be provided during training.

Method	Detector Training		LVIS
	Ext. Data	Pre-train Model	APr
ViLD [14]	-	CLIP (ViT-B/32)	16.3
OV-DETR [35]	-	CLIP (ViT-B/32)	17.4
RegionCLIP [40]	CC3M	CLIP (RN50x4)	22.0
CORA (Ours)	-	CLIP (RN50x4)	22.2
MEDet [7]	CC3M	CLIP (ViT-B/32)	22.4
Detic [42]	IN-21k	CLIP (text encoder)	26.2
CORA$^+$ (Ours)	IN-21k	CLIP (text encoder)	**28.1**

Table 2. Results on the LVIS [16] OVD benchmark.

Method	CLIP model	Novel	Base
CLIP	RN50	58.2	58.6
CLIP-Adapter [12]	RN50	63.0	80.6
CoOp [41]	RN50	64.4	75.7
CORA	RN50	65.1	70.0
CLIP	RN50x4	63.9	62.7
CORA	RN50x4	74.1	76.0

Table 3. Results on the region classification task evaluated in mAP. We compare our method with the original CLIP regional classifier and other baseline methods.

of-the-art RegionCLIP [40] by 2.4 AP50. When extra data is available, the performance can be further boosted to 43.1 AP50. The results on LVIS [16] OVD benchmark is shown in Tab. 2.

Note that among the baseline methods, VL-PLM [39], ViLD [14] and OV-DETR [35] use novel class names during training in order to recognize the potential novel objects and assign pseudo labels for them. Consequently, a new detector needs to be trained whenever there is a new set of categories to be detected. CORA can generalize to any combination of novel categories once trained without tedious re-training. Other compared methods relying on CLIP require image tag annotations extracted from language descriptions [7, 31, 40] or image labels [42] during training. We argue that the extra annotations do not provide additional information over CLIP. Instead, they serve as a media to transfer the knowledge from CLIP to the detector. Our method directly adapts the CLIP model to obtain region classifier, thus no extra image-text data is needed.

In this work, both region prompting and anchor pre-matching aim to generalize the knowledge learned from base classes to novel classes. Consequently, the performance gap between novel and base classes is significantly lowered than the compared methods. Note that in OVD, the performance is evaluated by the generalization capability towards novel classes, rather than the bases classes on which they are trained.

4.4. Effectiveness of Region Prompting

Region Prompting. Since object classification is decoupled from localization in this work, the CLIP-based region classifier can be directly evaluated by the region classification task. After trained on the base-class annotations of COCO dataset, we evaluate the CLIP-based region classifier on both base and novel classes in the validation set. We use mean average precision (mAP) as our evaluation metric.

Tab. 3 shows our main result. Directly evaluating on the CLIP model without further training already achieves considerable performance of 58.2 mAP on the novel classes. We compare our result with two competitive methods from the adapter and prompt tuning literatures. CLIP-Adapter [12] adopts an additional bottleneck layer to learn new features and performs residual style feature blending with the original pre-trained features. CoOp [41] prepends shared learnable prompts to the text embeddings before en-

CLIP model	Feature	Novel	Base
RN50	whole-image	43.8	40.5
RN50	regional	58.6	58.2

Table 4. Comparison of the whole-image classification pipeline and the regional classification pipeline. Results are reported in mAP.

Query Conditioning	Per-Class Post-Matching	CLIP-Aligned Labeling	Novel
✗	✗	✓	26.0
✓	✗	✓	29.9
✓	✓	✗	40.9
✓	✓	✓	**41.7**

Table 5. Ablation studies on anchor pre-matching and CLIP-aligned labeling. Anchor pre-matching consists of query conditioning and per-class post-matching.

coded by the CLIP text encoder. Experiments show that the compared methods bias strongly towards base classes. Note that the baseline methods adapts the CLIP model by tuning the text input or output feature, which are irrelevant or distant to the regional feature, where the mismatch between regional and whole-image feature occurs. On the contrary, region prompting directly prompts the mismatched features, thus generalizes better to the novel classes, achieving a performance gain of 6.9 mAP over CLIP on the novel classes. Region prompting also scales with larger backbones. On the RN50x4 CLIP backbone, region prompting further boosts CLIP by 10.2% mAP over the corresponding CLIP model.

Comparison with whole-image classification. A common practice of region classification by CLIP is to crop the regions and classify them as separate images. We compare the region classifier with the common practice on the original CLIP weights without region prompting. As shown in Tab. 4, when classifying a region as a whole-image, the performance is significantly lower than using the regional feature, despite the extra computation. We attribute the performance gap to the missing context of cropped images.

4.5. Effectiveness of Anchor Pre-Matching

We conduct ablation studies to validate anchor pre-matching and the proposed training techniques.

Anchor Pre-Matching. Anchor pre-matching consists of two separate operations: query conditioning and per-class post-matching. We analyze their effects in Tab. 5. Firstly, we train a model without anchor pre-matching. We keep using DAB-DETR [26] as the localizer, and use the same number of object queries and anchor boxes for fair comparison. The object queries are not pre-matched to classes, and are initialized as 0 as done in DAB-DETR. Since the model predictions are not pre-matched, the per-class post-matching is replaced by the vanilla one-to-one matching mechanism in DETR. Then, we classify the anchor boxes and condition the object queries on the class name embeddings, but do not set constraint on the post-matching, which boosts the performance on novel classes by 3.9 AP50. After adopting the full anchor pre-matching, the performance is significantly boosted by 10.8 AP50.

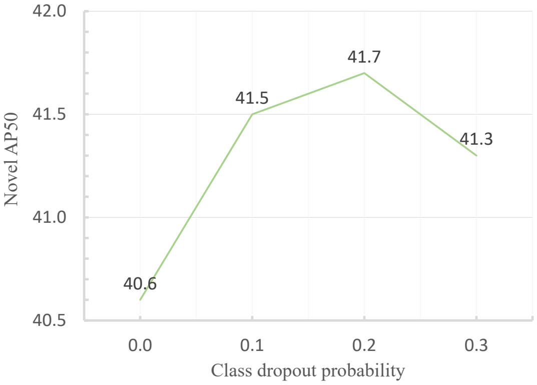


Figure 4. Different choices of p for class dropout.

Training Techniques. Class dropout and CLIP-aligned labeling are two training techniques that help the model generalize better. In Fig. 4, we examine the effect of different dropout probability. We find that the models trained with class dropout consistently outperform the baseline, and $p = 0.2$ gives the best performance. In Tab. 5, we validate the effectiveness of CLIP-Aligned Labeling.

5. Conclusion

The core challenge in open-vocabulary detection is how to effectively transfer the knowledge learned from the base classes to the unseen novel classes for evaluation. In this work, we directly adapt the CLIP into a region classifier, and mitigate the distribution gap between whole-image feature and regional feature through region prompting, which successfully generalizes to the novel categories. Different from prior works that rely on a fixed RPN for novel class localization, we achieve efficient class-aware localization by the proposed anchor pre-matching mechanism. Experiments show that our method can better transfer the knowledge from base classes to unseen novel classes with a smaller gap than prior works. We hope our work can help other researchers gain better insight on the OVD problem and develop better open-vocabulary detectors.

Acknowledgement. This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK. This project is also supported by SenseTime Collaborative Research Grant.

References

[1] Deformable detr: Deformable transformers. 2020. 2

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274, 2022. 2

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. ArXiv, abs/2203.17274, 2022. 2

[4] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. ArXiv, abs/1804.04340, 2018. 6

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. ArXiv, abs/2005.14165, 2020. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. ArXiv, abs/2005.12872, 2020. 1, 2

[7] 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. 1, 7

[8] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. ArXiv, abs/2207.13085, 2022. 2

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. ArXiv, abs/1504.00325, 2015. 7

[10] 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. 2022. 2

[11] 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

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. ArXiv, abs/2110.04544, 2021. 7

[13] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3601–3610, 2021. 2

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In ICLR, 2022. 1, 7

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

[16] 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, 6, 7

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017. 1

[18] Ding Jia, Yuhui Yuan, Haodi He, Xiao pei Wu, Haojun Yu, Weihong Lin, Lei huan Sun, Chao Zhang, and Hanhua Hu. Detrs with hybrid matching. ArXiv, abs/2207.13080, 2022. 2

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser Nam Lim. Visual prompt tuning. ArXiv, abs/2203.12119, 2022. 2

[20] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. ArXiv, abs/2112.04478, 2021. 2

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10955–10965, 2022. 2

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), abs/2101.00190, 2021. 2

[23] 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. 5

[24] 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. 2, 6

[25] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. arXiv preprint arXiv:2208.03550, 2022. 2

[26] Shilong Liu, Feng Li, Hao Zhang, Xiao Bin Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. ArXiv, abs/2201.12329, 2022. 2, 5, 6, 8

[27] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. CoRR, abs/2110.07602, 2021. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019. 6

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In 17th European Conference on Computer Vision (ECCV). Springer, 2022. 1

[30] 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 ICML, 2021. 1, 6

[31] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. ArXiv, abs/2207.03482, 2022. 1, 7

[32] 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. 4

[33] 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. 7

[34] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. CoRR, abs/2208.08340, 2022. 2

[35] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. ArXiv, abs/2203.11876, 2022. 1, 2, 5, 7

[36] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14388–14397, 2021. 1, 2, 7

[37] 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

[38] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating DETR convergence via semantic-aligned matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 949–958, 2022. 2, 6

[39] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar BG, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N. Metaxas. Exploiting unlabeled data with vision and language models for object detection. CoRR, abs/2207.08954, 2022. 1, 7

[40] 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. 1, 2, 6, 7

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. Int. J. Comput. Vis., 130:2337–2348, 2022. 7

[42] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In ECCV, 2022. 7

[43] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. In arXiv preprint arXiv:2103.07461, 2021. 6