Learning Background Prompts to Discover Implicit Knowledge for Open Vocabulary Object Detection

Jiaming Li[1]   Jiacheng Zhang[1]   Jichang Li[1,2]   Ge Li[3]   Si Liu[4]
Liang Lin[1]   Guanbin Li [1,5,6*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Science, The University of Hong Kong, Hong Kong
[3]SECE, Shenzhen Graduate School, Peking University, Shenzhen, China
[4]Institute of Artificial Intelligence, Beihang University, China
[5]GuangDong Province Key Laboratory of Information Security Technology
[6]Research Institute, Sun Yat-sen University, Shenzhen, China

{lijm48,zhangjch58}@mail2.sysu.edu.cn, {liguanbin,linlng}@mail.sysu.edu.cn
csjcli@connect.hku.hk, lige@pku.edu.cn, liusi@buaa.edu.cn

Abstract

Open vocabulary object detection (OVD) aims at seeking an optimal object detector capable of recognizing objects from both base and novel categories. Recent advances leverage knowledge distillation to transfer insightful knowledge from pre-trained large-scale vision-language models to the task of object detection, significantly generalizing the powerful capabilities of the detector to identify more unknown object categories. However, these methods face significant challenges in background interpretation and model overfitting and thus often result in the loss of crucial background knowledge, giving rise to sub-optimal inference performance of the detector. To mitigate these issues, we present a novel OVD framework termed LBP to propose learning background prompts to harness explored implicit background knowledge, thus enhancing the detection performance w.r.t. base and novel categories. Specifically, we devise three modules: Background Category-specific Prompt, Background Object Discovery, and Inference Probability Rectification, to empower the detector to discover, represent, and leverage implicit object knowledge explored from background proposals. Evaluation on two benchmark datasets, OV-COCO and OV-LVIS, demonstrates the superiority of our proposed method over existing state-of-the-art approaches in handling the OVD tasks.
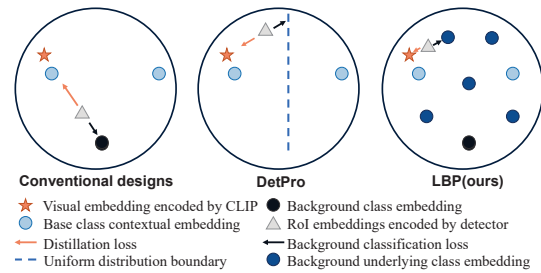
Figure 1. An example to illustrate previous and our designs in background interpretation. Conventional designs use a single background embedding to push the RoI embedding away from the CLIP embedding. DetPro [4] proposes to uniformly push the RoI embedding away when the CLIP embedding nears a base class embedding, leading to a loss of class relation. Our LBP, on the other hand, learns multiple background underlying class embeddings, effectively preserving class relations and alleviating loss conflict. "Distillation loss" uses knowledge distillation to align visual features encoded by the decoder with CLIP embeddings, while "Background classification loss" refers to the classification loss for background proposals.

1. Introduction

Compared to conventional vision tasks [11, 12, 18–22, 39], object detection has witnessed significant success in research, such as [26, 30, 32, 41], however, solely detecting and classifying objects within known categories (base classes) during inference significantly diminishes its generalization capacity in real-world applications. Open Vo-

---

*Corresponding author.

cabulary Object Detection (OVD) emerges as a prospective means to overcome this restriction, endowed with the capability to detect unseen categories (novel classes), without the explicit need for annotations.

Leveraging large-scale pre-trained Vision and Language Models (PVLMs), exemplified by CLIP [29] and ALIGN [13], recent advances in OVD, e.g. [4, 7, 35], employ knowledge distillation to transfer insightful knowledge of PVLMs to the task of object detection, generalizing its powerful capabilities to detect more unknown object categories. However, these methods face a critical limitation in background interpretation. Specifically, these approaches tend to represent background proposals with a single "background class", encompassing all categories beyond foreground base classes. With solely a singular learnable embedding to interpret this background class, the trained model may fail to capture the diverse implicit knowledge within background proposals. This leads to the loss of essential background information, resulting in incomplete and ambiguous representations of unseen categories by the model. Consequently, the model is incapable of distinguishing objects from those unknown categories using their representations.

To tackle this problem, researchers experimented with various strategies like binary cross-entropy loss [2] or soft background loss [4] for background interpretation. However, their basic assumption that background proposals are uniformly dissimilar to any foreground category overlooks the nuanced class relations between foreground and background classes, as illustrated in Figure 1. Additionally, these methods struggle with model overfitting during training due to an abundance of fully supervised data from base classes, which makes detectors biased toward those categories. Recent advances like [2, 5, 6, 40] propose mining implicit objects of novel categories from background proposals and employing pseudo-labeling to enhance their interpretation. Yet, these methods often rely on additional prior knowledge, such as the names of novel categories, leaving substantial unexplored knowledge.

In this paper, we present LBP, a novel framework for open-vocabulary object detection. Without any prior knowledge, LBP proposes learning background prompts to harness explored implicit background knowledge, thus enhancing the detection performance w.r.t. base and novel categories. Specifically, the LBP framework initially introduces a Background Category-specific Prompt module. It discovers and represents background underlying categories estimated from background proposals by leveraging learnable category-specific contexts, consequently resulting in improved background interpretation. Then, an online module, namely Background Object Discovery, is introduced to further explore and exploit implicit object knowledge correlated with those estimated underlying categories from

background proposals, significantly contributing to mitigating model overfitting. Moreover, an Inference Probability Rectification module is presented to address conceptual overlaps between estimated background categories and novel categories provided during inference. This rectification enables the model to accurately compute probabilities for novel categories, thereby significantly enhancing the detector performance. The contributions are summarized as follows.

- We propose a novel framework, termed LBP, for open-vocabulary object detection, where learning background prompts is presented to harness explored implicit background knowledge, thereby enhancing detection of both base and novel categories during inference.
- We devise three modules: Background Category-specific Prompt, Background Object Discovery, and Inference Probability Rectification, to empower the detector to discover, represent, and leverage implicit object knowledge explored from background proposals.
- Evaluation on two benchmark datasets, OV-COCO [25] and OV-LVIS [9], demonstrates the superiority of our proposed method over existing state-of-the-art approaches in handling the OVD tasks.

2. Related Work

Open vocabulary object detection (OVD). OVR-CNN [38] pioneered open vocabulary object detection, by expanding object detection capabilities through image-caption datasets. Subsequently, prevailing approaches to tackle OVD have primarily relied on fine-tuning PVLMs [16, 17, 23, 24, 28, 42]. However, retraining PVLMs with large-scale learnable model weights is resource-intensive. Consequently, more recent approaches rooted in knowledge distillation [27, 34, 36] to address OVD by leveraging knowledge distilled from CLIP [29] into the detector, predominantly adopting source-free strategies. For example, ViLD [7] aligns detector-encoded features with CLIP embeddings, while BARON [35] generates region ensembles using random masking within grid spaces to distill visual concept co-occurrence within a scene. Despite partially transferring knowledge of novel categories from PVLMs, detectors exhibit bias towards base classes due to disparities between vision-language alignment and detection tasks. To mitigate this bias, researchers explore pseudo labeling techniques utilizing prior knowledge of novel category names [40] or weakly-supervised annotations to the detection dataset [2, 6, 46]. However, integrating additional knowledge poses limitations in real-world scenarios.

Our proposed method extracts background implicit knowledge about classes beyond known/base categories from the provided detection dataset without requiring prior information about novel categories during inference. This
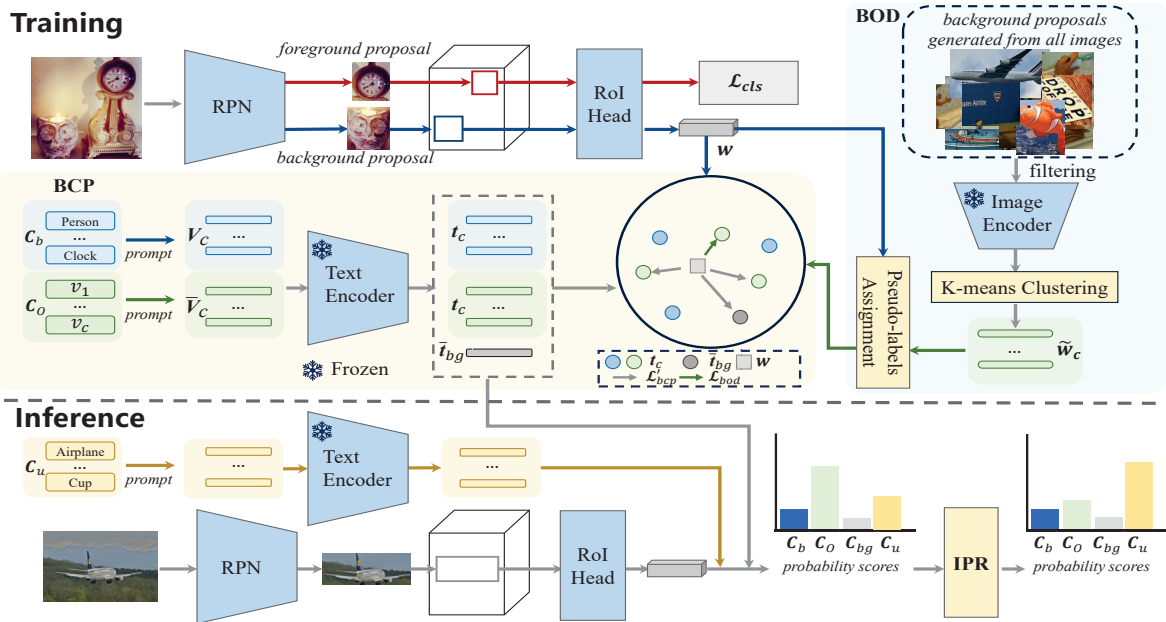
Figure 2. An overview of the proposed LBP approach, consisting of three modules: Background Category-specific Prompt (BCP), Background Object Discovery (BOD), and Inference Probability Rectification (IPR). During training, BCP is first proposed to discover and represent background underlying categories, estimated from background proposals, with learnable background category-specific contexts. Then, BOD is presented to employ k-means clustering on background proposals across all images to harness implicit objects explored from background underlying knowledge. During inference, IPR is introduced to rectify probability scores of novel categories provided, by loosening their conceptual overlaps with background underlying categories estimated from background proposals.

utilization of implicit knowledge enables the detector to identify objects from unknown categories, enhancing its applicability across diverse scenarios.

Prompt tuning for OVD: Prompt tuning, stemming from natural language processing, has evolved with PVLM advancements to enhance their efficacy in specific downstream tasks [15]. CoOp, proposed in [44], introduces learnable contexts for 2D vision-language classification. CoCoOp, an extension proposed by [43], further enhances CoOp by generating input-conditioned tokens for each image. Various visual prompting approaches advocate integrating prompts into the image encoder of PVLMs, including works such as [1, 3, 8, 14].

In the realm of OVD, DetPro [29] pioneered prompt engineering by implementing a soft background loss, significantly improving the detector's ability to represent specific categories. Similarly, proposals for pre-training prompts on large-scale datasets [5, 31] aim to generate more universally applicable prompts. In contrast to DetPro [29], our approach has no reliance on uniform distribution assumption. Instead, by prompting underlying background categories estimated from proposals, our method enhances the learning of unknown category representations, leading to significant performance gains in detection.

3. Approach

3.1. Preliminaries

Problem formulation. Open vocabulary object detection (OVD) aims at seeking an optimal object detector capable of recognizing objects from both base and novel (previously unseen) categories, denoted by $\mathcal{C}_b = \{c_i\}_{i=1}^{n_b}$ and $\mathcal{C}_u = \{c_i\}_{i=1}^{n_u}$, respectively, in the inference dataset $\mathcal{D}_I$. This detector is achieved through optimization on the training dataset $\mathcal{D}_T$, annotated exclusively with instances from the base classes within $\mathcal{C}_b$. Here, $\mathcal{C}_b \cap \mathcal{C}_u = \emptyset$, while n_b and n_u represent the cardinality of the respective category sets.

Built upon Faster R-CNN [30], given an input image, it is initially encoded by Faster R-CNN into global image features and generates a set of proposals through a region proposal network (RPN). In general, these region proposals are categorized into *foreground* proposals and *background* proposals [4], designated as $\mathcal{P}$ and $\mathcal{N}$, respectively. The foreground proposals in $\mathcal{P}$ encompass all base categories within $\mathcal{C}_b$, while all background proposals from $\mathcal{N}$ are collectively classified as a singular **super-class** (denoted as c_{bg} here), representing any category outside of $\mathcal{C}_b$.

Knowledge distillation. Leveraging large-scale pre-trained large-scale Vision and Language Models (PVLMs), specif-

ically CLIP [44], knowledge distillation empowers Faster R-CNN to effectively detect objects from arbitrary vocabularies by transferring insightful knowledge from CLIP. This process involves aligning the proposal features generated by the detector (actually output by the RoI Head of the detector, denoted by $\boldsymbol{w}(\cdot)$) with those extracted by the image encoder of CLIP (denoted by $\mathcal{I}(\cdot)$), effectively simulating coherence between the feature spaces of the detector and CLIP. Consequently, this process implicitly projects the visual embedding space of the detector into the textual embedding space of CLIP (output by the text encoder of CLIP, denoted by $\mathcal{T}(\cdot)$). In this way, the original classifier of Faster R-CNN is replaced.

Foreground/background interpretation. Given class c from $\mathcal{C}_b$, its contextual prompt is defined as 'a photo of {category} in the scene', noted as $\boldsymbol{V}_c$. Here, '{category}' denotes the contextual word for class c. Then its contextual embedding is obtained by feeding $\boldsymbol{V}_c$ into the CLIP text encoder $\mathcal{T}(\cdot)$, as follows,

$$\boldsymbol{t}_c = \mathcal{T}(\boldsymbol{V}_c). \quad (1)$$

However, it is challenging to determine the contextual word for the "*background*" class $\boldsymbol{c}_{bg}$. ViLD [7] or BARON [35] represents it with a learnable contextual embedding, without encoding it via CLIP's text encoder $\mathcal{T}(\cdot)$, thus defining it as $\boldsymbol{t}_{bg}$.

Probability calculation. Having obtained $\{\boldsymbol{t}_c | c \in \mathcal{C}_b\} \cup \{\boldsymbol{t}_{bg}\}$, let $\boldsymbol{w}(\cdot)$ denote the visual embedding of a proposal generated by the RoI Head of the object detector. Consider $\mathcal{C}_b^{bg} = \mathcal{C}_b \cup \{\boldsymbol{c}_{bg}\}$. For any proposal $x \in \mathcal{P} \cup \mathcal{N}$ from $\mathcal{D}_T$ with its visual embedding $\boldsymbol{w}(x)$, the probability of categorizing this proposal under the category $c \in \mathcal{C}_b^{bg}$ is defined as follows:

$$p(c|x; \mathcal{C}_b^{bg}) = \frac{\exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_c)/\tau)}{\sum_{c' \in \mathcal{C}_b^{bg}} \exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_{c'})/\tau)}, \quad (2)$$

where $\exp(\cdot)$ and $\cos(\cdot)$ represent the exponential function and cosine similarity, respectively, while τ is the temperature parameter used for rescaling the values.

During inference, once the novel classes from $\mathcal{C}_u$ are introduced, the probability of a proposal from $\mathcal{D}_I$ being classified as class c of $\mathcal{C}_b \cup \{\boldsymbol{c}_{bg}\} \cup \mathcal{C}_u$ during detection can be computed simply by replacing $\mathcal{C}_b^{bg}$ in Eq. (2) with $\mathcal{C}_{b,u}^{bg} = \mathcal{C}_b \cup \{\boldsymbol{c}_{bg}\} \cup \mathcal{C}_u$.

3.2. The Proposed Method

In this paper, we introduce LBP, a novel framework designed for open-vocabulary object detection. The key idea of LBP is that without any prior knowledge, learning background prompts is proposed to harness explored implicit background knowledge. This ensures the model attains improved background interpretation and decreased model overfitting, thus empowering the detector capable of better recognizing both base and novel categories. Specifically, the LBP framework initially introduces a Background Category-specific Prompt (**BCP**) module, discovering underlying categories from background proposals, and representing those using learnable context-specific prompts. Moreover, we introduce an online Background Object Discovery (**BOD**) to further exploit implicit object knowledge w.r.t. those estimated underlying categories, consequently alleviating model bias towards base categories, and present an Inference Probability Rectification (**IPR**) module to resolve conceptual overlaps between estimated background categories and novel categories during inference, leading to precisely computed probability scores for novel categories. Our primary focus lies in exploring background interpretation, distinct from prevailing OVD approaches rooted in knowledge distillation. This enables seamless integration of our method into existing OVD frameworks [35] and ViLD [7], etc. An overview of the proposed approach is illustrated in Figure 2.

3.2.1 Background Category-specific Prompt

In this section, we model the underlying categories for background proposals in OVD and learn the corresponding category-specific prompts. However, the absence of prior knowledge of those categories necessitates the estimation of their optimal number in background proposals.

Towards this end, we first produce background proposals for all training images using a class-agnostic RPN trained on $\mathcal{D}_T$ with only base classes, adopting a similar technique as in VL-PLM [40]. Those proposals are subsequently filtered based on RPN scores (with a threshold θ), and Non-Maximum Suppression (NMS). Then, we obtain features for these filtered proposals generated from the image encoder of CLIP, and we refer to [10, 41] to estimate the optimal number of the underlying categories in background proposals via K-mean clustering [33], denoted by n_o, in the obtained proposal features. During this procedure, potential ambiguity in data partitioning during estimation may arise due to feature overlap, leading to the underestimation of n_o. Consequently, we expand the count of background underlying categories to $n_o + n_a$, where n_a serves as a pre-defined hyper-parameter. Here, let $\mathcal{C}_O = \{c_i\}_{i=1}^{n_o+n_a}$ denote those estimated background categories, for simplicity.

However, due to the absence of prior knowledge of $\mathcal{C}_O$, identifying their contextual words is a non-trivial task. To address this issue, we propose to learn category-specific context vectors. In this case, the contextual words of those background underlying categories can be described with the learned continuous vectors. Specifically, we adopt the prompt form for a given class $c \in \mathcal{C}_O$ as follows:

$$\overline{\boldsymbol{V}}_c = \text{'a photo of } \{v_c\} \text{ in the scene'}, \quad (3)$$

where the context vectors $\{v_c | c \in \mathcal{C}_O\}$ are learnable. Following Eq. (1), the corresponding contextual embeddings can be acquired as $\bar{t}_c = \mathcal{T}(\overline{V}_c)$.

Additionally, we present a specific **sub-class** category, $\bar{c}_{bg}$, designed to represent the "background" within background proposals, distinct from the super-class "*background*" c_{bg}. Specifically, $\bar{c}_{bg}$ covers unexplored categories not estimated in $\mathcal{C}_O$ and includes any **untargeted** object within background proposals. Similar to t_{bg}, we define a learnable contextual embedding, $\bar{t}_{bg}$, for this category $\bar{c}_{bg}$.

Once we obtain $\{t_c | c \in \mathcal{C}_b\} \cup \{\bar{t}_c | c \in \mathcal{C}_O\} \cup \{\bar{t}_{bg}\}$, the corresponding set of categories can be defined as $\mathcal{C}_{b,o}^{bg} = \mathcal{C}_b \cup \mathcal{C}_O \cup \{\bar{c}_{bg}\}$. With reference to Eq. (2), for any proposal $x \in \mathcal{P} \cup \mathcal{N}$, its probability of being classified as class $c \in \mathcal{C}_{b,o}^{bg})$ can be defined as $p(c|x; \mathcal{C}_{b,o}^{bg})$. At this point, the model optimization in the **BCP** for the proposals in $\mathcal{P}$ and $\mathcal{N}$ can be achieved using the following cross-entropy losses:

$$\mathcal{L}_{cls} = \frac{1}{|\mathcal{P}|} \sum_{x \in \mathcal{P}} - \log p(c = \boldsymbol{C}(x)|x; \mathcal{C}_{b,o}^{bg}), \quad (4)$$

$$\mathcal{L}_{bcp} = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} - \log \sum_{c \in \mathcal{C}_O \cup \{\bar{c}_{bg}\}} p(c|x; \mathcal{C}_{b,o}^{bg}), \quad (5)$$

where $\boldsymbol{C}(x)$ represents the true label class of the given proposal x. $\mathcal{L}_{bcp}$ enhances the sum of probabilities of all background underlying categories estimated and $\bar{c}_{bg}$ within each background proposal, enabling the detector to softly learn the probabilities scores assigned for those categories.

However, contextual embeddings of $\mathcal{C}_O \cup \{\bar{c}_{bg}\}$ might not be sufficient to represent all background proposals with diverse scenes. When a background proposal belongs to $\bar{c}_{bg}$ and cannot be adequately represented by its contextual embedding, the visual embeddings of that proposal will be distant from any contextual embeddings of the estimated categories from $\mathcal{C}_O$. Consequently, the sum of probabilities of those categories, i.e., $\sum_{c \in \mathcal{C}_O \cup \{\bar{c}_{bg}\}} p(c|x; \mathcal{C}_{b,o}^{bg})$ in Eq. (5), denoted as $\boldsymbol{p}_o^{bg}$, for simplicity, would become very small.

In this situation, we should uniformly push the visual embeddings of that proposal towards the estimated background underlying categories and $\bar{c}_{bg}$, given that the proposal is distinct from those categories. When $\boldsymbol{p}_o^{bg} < \gamma$, we introduce an additional softer background loss component to relax $\mathcal{L}_{bcp}$ for all background proposals in $\mathcal{N}$ as follows,

$$\mathcal{L}_{rlx} = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} \frac{1}{n_{oa}} \sum_{c \in \mathcal{C}_O \cup \{\bar{c}_{bg}\}} - \log p(c|x; \mathcal{C}_{b,o}^{bg}), \quad (6)$$

where $n_{oa} = n_o + n_a + 1$ represents the size of $\mathcal{C}_O \cup \{\bar{c}_{bg}\}$, while γ is a threshold with a small value. Therefore, the final version of the loss component in this **BCP** for background proposals to optimize the model is formulated as,

$$\mathcal{L}'_{bcp} = \begin{cases} \mathcal{L}_{bcp}, & \text{if } \boldsymbol{p}_o^{bg} \geq \gamma, \\ \mathcal{L}_{rlx}, & \text{otherwise.} \end{cases} \quad (7)$$

Note that $\mathcal{L}'_{bcp}$ enriches the discrimination capability of the detector by incorporating diverse contextual embeddings of estimated background categories, diverging from the soft background loss in DetPro [4]. This enables the model to exploit implicit object knowledge w.r.t. those estimated underlying categories, while also reconciling training conflicts between the losses utilized for classification and distillation, as highlighted in prior works [35].

3.2.2 Background Object Discovery

To enhance model training, we introduce an online Background Object Discovery (**BOD**) module to effectively discover and exploit unseen objects. This aims to extract implicit objects w.r.t. the underlying categories estimated from background proposals. To simplify, we here decompose $\mathcal{C}_O$ into $\mathcal{C}'_O = \{c_i\}_{i=1}^{n_o}$ and $\mathcal{C}_a = \{c_i\}_{i=1}^{n_a}$ in this section.

During initial training, with a given n_o, we conduct k-means clustering (where k is set to n_o) on the set of visual embeddings $\{\mathcal{I}(x) | x \in \mathcal{N}\}$ generated by the image encoder of CLIP for background proposals as in Sec. 3.2.1. Then, we subsequently obtain the embeddings of the cluster centers, denoted as $\{\tilde{\boldsymbol{w}}_c | c \in \mathcal{C}'_O\}$. These cluster centers function as the embedding centers for the estimated background categories within $\mathcal{C}'_O$, allowing for online generation of pseudo labels from background proposals in each training batch. Throughout the training process, motivated by VL-PLM proposed in [40], background proposals of each training batch undergo being filtered based on RPN scores using a threshold of θ, with an additional step to filter out proposals overlapping with ground-truth boxes. Subsequently, CLIP is employed to generate pseudo labels, preventing the detector from being biased towards estimated background categories.

Then, for a proposal $x \in \mathcal{N}$, its probability score of being classified as $c \in \mathcal{C}'_O$ is calculated as follows,

$$\tilde{p}(c|x; \mathcal{C}'_O) = \frac{\exp(\cos(\mathcal{I}(x), \tilde{\boldsymbol{w}}_c)/\tau)}{\sum_{c' \in \mathcal{C}'_O} \exp(\cos(\mathcal{I}(x).\tilde{\boldsymbol{w}}_{c'})/\tau)}. \quad (8)$$

Once such background proposal has probability scores corresponding to all categories in $\mathcal{C}'_O$, we choose the predicted class label with the highest score as its pseudo label,

$$\boldsymbol{y}^o(x) = \operatorname*{argmax}_c \{\tilde{p}(c|x; \mathcal{C}'_O) | c \in \mathcal{C}'_O\}. \quad (9)$$

To eliminate unconfident pseudo labels generated above, we will filter proposals based on probability scores of pseudo labels using a threshold θ. Referring to VL-PLM [40], we also apply per-class NMS and use RoI Head to refine their box predictions, generating final pseudo labels. Afterward, those final pseudo labels are used to assign class labels of $\mathcal{C}'_O$ for all background proposals $\mathcal{N}^B \subseteq \mathcal{N}$ in each training batch. Let $\mathcal{N}_p^B$ denote the set of positive

background proposals assigned class labels from $\mathcal{C}'_O$, while the remaining proposals are all collected as $\mathcal{N}^B_n$.

Therefore, the loss component proposed by this **BOD** module to optimize the model is formulated as follows,

$$\mathcal{L}_{bod} = \frac{1}{|\mathcal{N}^B_p|} \sum_{x \in \mathcal{N}^B_p} - \log p(c = \boldsymbol{y}^o(x)|x; \mathcal{C}^{bg}_{b,o})$$
$$+ \lambda_{bg} \frac{1}{|\mathcal{N}^B_n|} \sum_{x \in \mathcal{N}^B_n} - \log \sum_{c \in \mathcal{C}_a \cup \{\bar{\boldsymbol{c}}_{bg}\}} p(c|x; \mathcal{C}^{bg}_{b,o}), \tag{10}$$

where λ_{bg} is a loss weight with a small value used for background proposals. This loss component emphasizes aligning visual embeddings of background proposals with their contextual embeddings related to estimated background categories from $\mathcal{C}_O$. It is especially critical for majority classes within $\mathcal{C}_O$, which are more prone to knowledge loss. Additionally, it empowers **BOD** to leverage insights from implicit objects within background proposals.

Training Objective. The final training objective for model optimization is a combination of the loss components formulated by Eq. (4), Eq. (7), and Eq. (10), respectively:

$$\mathcal{L}_{final} = \mathcal{L}_{cls} + \mathcal{L}'_{bcp} + \mathcal{L}_{bod}. \tag{11}$$

It's worth noting that, the text encoder and image encoder of CLIP should be frozen during model training.

3.2.3 Inference Probability Rectification

Once performed **BCP** and **BOD**, the detector capability of recognizing previously unseen classes has been significantly enhanced. However, a new challenge arises during inference. The background underlying categories $\mathcal{C}_O$, estimated from background proposals during training, might share semantics similarities with the novel classes $\mathcal{C}_u$ of the detector aiming to classify during inference, vividly symbolized by $\mathcal{C}_O \cap \mathcal{C}_u \neq \emptyset$. In such cases, there exist conceptual overlaps of contextual embeddings between those two types of categories during inference. This overlap could hinder the accurate computation of probability scores for novel classes, leading to detection ambiguity during inference. To overcome this, an Inference Probability Rectification (**IPR**) is presented to enable the detector to precisely predict the probabilities of novel classes during inference.

To be specific, let $\mathcal{C}^{bg}_{b,u,o} = \mathcal{C}_b \cup \mathcal{C}_u \cup \mathcal{C}_O \cup \{\bar{\boldsymbol{c}}_{bg}\}$. Then, with reference to Eq. (2), the probability of a proposal $x \in \mathcal{D}_I$ being classified as class $c \in \mathcal{C}_b \cup \mathcal{C}_u$ during inference can be computed as follows,

$$p(c|x; \mathcal{C}^{bg}_{b,u,o}) = \frac{\exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_c)/\tau)}{\boldsymbol{\Sigma}^{b,u} + \boldsymbol{\Sigma}^o + \boldsymbol{\Sigma}^{bg}}, \tag{12}$$

where

$$\boldsymbol{\Sigma}^{b,u} = \sum_{c' \in \mathcal{C}_b \cup \mathcal{C}_u} \exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_{c'})/\tau),$$
$$\boldsymbol{\Sigma}^o = \sum_{c' \in \mathcal{C}_O} \exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_{c'})/\tau),$$
$$\boldsymbol{\Sigma}^{bg} = \sum_{c' \in \{\bar{\boldsymbol{c}}_{bg}\}} \exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_{c'})/\tau). \tag{13}$$

Assuming that a proposal x carries its true class label $c \in \mathcal{C}_u$ simultaneously sharing conceptual overlap with categories in $\mathcal{C}_O$, then the probability computation of for this category, namely $p(c|x; \mathcal{C}^{bg}_{b,u,o})$, tends to underestimate its value. This underestimation occurs due to the presence of c simultaneously contributing in both $\boldsymbol{\Sigma}^{b,u}$ and $\boldsymbol{\Sigma}^o$ in the denominator of Eq. (12). In essence, $p(c|x; \mathcal{C}^{bg}_{b,u,o})$ becomes smaller than its true probability value, denoted as $P(c|x)$, which is a theoretical but unknown value, being discussed later. To address this issue, our goal is to resolve the conceptual overlaps between novel categories in $\mathcal{C}_u$ and estimated background underlying categories within $\mathcal{C}_O$. This aims to reduce the influence of c contributing $\boldsymbol{\Sigma}^o$, leading to a re-estimation of $\boldsymbol{\Sigma}^o$ and thus resulting in $\tilde{\boldsymbol{\Sigma}}^o$.

Here, we simplify the function $s(\cdot, \cdot) = \exp(\cos(\cdot, \cdot)/\tau)$ in Eqs. (12) to (13) and term its resulting value as the cosine exponential score. As per the definitions of the softmax function, individual probabilities for each category are computed by normalizing the cosine exponential scores, dividing each by the total sum of cosine exponential scores corresponding to all categories. We hypothesize that the cosine exponential score proportionally reflects its true probability. Hence, $\tilde{\boldsymbol{\Sigma}}^o$ can be represented as:

$$\tilde{\boldsymbol{\Sigma}}^o = \sum_{c' \in \mathcal{C}_O} s(\boldsymbol{w}(x), \boldsymbol{t}_{c'}) \cdot (1 - \frac{\sum_{c'' \in \mathcal{C}_u} P(c', c''|x)}{P(c'|x)}). \tag{14}$$

According to the Multiplication Rule of Joint Probabilities, the joint probability $P(c', c''|x)$ for any $c' \in \mathcal{C}_O$ and $c'' \in \mathcal{C}_u$ can be estimated as follows:

$$P(c', c''|x) = P(c'|x)P(c''|x, c'). \tag{15}$$

With Eq. (15), Eq. (14) can be reformulated as follows:

$$\tilde{\boldsymbol{\Sigma}}^o = \sum_{c' \in \mathcal{C}_O} s(\boldsymbol{w}(x), \boldsymbol{t}_{c'}) \cdot (1 - \sum_{c'' \in \mathcal{C}_u} P(c''|x, c')). \tag{16}$$

However, estimating $P(c''|x, c')$ poses challenges. Hence, we assume $P(c''|x, c')$ to be sample-agnostic, resulting in $P(c''|c')$. Given that the embedding space serves as a metric space for probability calculation, $P(c''|c')$ can be estimated by leveraging the cosine similarity between the contextual embeddings of these two categories as follows,

$$P(c''|x, c') = P(c''|c') = \frac{s(\boldsymbol{t}_{c'}, \boldsymbol{t}_{c''})}{\sum_{c \in \mathcal{C}^{bg}_{b,u,o} \setminus \{c'\}} s(\boldsymbol{t}_{c'}, \boldsymbol{t}_c)}. \tag{17}$$

Considering Eq. (16) and Eq. (17) into Eq. (12), the probability of a proposal $x \in \mathcal{D}_I$ being classified as class $c \in \mathcal{C}_b \cup \mathcal{C}_u$ during inference can be reformulated as follows,

$$p(c|x; \mathcal{C}_{b,u,o}^{bg}) = \frac{\exp(\cos(\boldsymbol{w}(x), \boldsymbol{t}_c)/\tau)}{\boldsymbol{\Sigma}^{b,u} + \tilde{\boldsymbol{\Sigma}}^{o} + \boldsymbol{\Sigma}^{bg}}. \quad (18)$$

4. Experiments

4.1. Experimental Setups

Datasets. To assess the effectiveness of the proposed LBP framework in handling the OVD task, we conducted experiments on two established object detection benchmark datasets: MS-COCO [25] and LVIS [9]. These experiments were conducted in traditional open vocabulary settings [7, 34, 35], referred to as OV-COCO and OV-LVIS, respectively. As outlined in prior works [34, 35], we divided 48 categories as base classes and 17 categories as novel classes in the OV-COCO task. The primary metric used to evaluate the detection performance is the mean average precision at IoU with a threshold of 0.50 (denoted as AP_{50}). Specifically, we represent AP_{50} for base and novel categories as AP_{50}^n and AP_{50}^b, respectively. For OV-LVIS, following [34, 35], we classified 337 rare categories as novel classes, while considering the remaining common and frequent categories as base classes (resulting in a total of 866 categories). Here, the detector's performance is evaluated using the mean average precision averaged across IoUs from 0.50 to 0.95 (denoted as AP). We report AP values for rare categories (AP_r), common categories (AP_c), frequent categories (AP_f), and all classes. Furthermore, we include experimental results for instance segmentation on LVIS. The metrics AP_{50}^n and AP_r are used as the primary measures to assess the detector's performance on OV-COCO and OV-LVIS, respectively.

Baselines. We compare LBP with the following state-of-the-art (SOTA) algorithms to handle the OVD task on OV-COCO: Detic [46], Object-centric-OVD [2], OV-DETR [37], RegionCLIP [42], ViLD [7], OADP [34] and BARON [35]. Besides, we perform our comparison on OV-LVIS with previous methods, including ViLD [7], DetPro [4] and BARON [35].

4.2. Comparisons with State-of-the-Arts

Results on OV-COCO and OV-LVIS are reported in Table 1 and Table 2, respectively. As shown, our LBP method outperforms previous state-of-the-art (SOTA) algorithms for the OVD task in all cases, validating its effectiveness.

Results on OV-COCO. Similar to OADP [34], we categorize the comparison baselines into four OVD benchmark settings: Vanilla OVD (V-OVD), Caption-based OVD (C-OVD), Generalized OVD (G-OVD), and Weakly Supervised OVD (WS-OVD). The emphasis of our proposed ap-

Method	Benchmark	Detector	AP_{50}^n	AP_{50}^b	AP_{50}
Detic [46]	WS-OVD	CenterNet2[45]	`27.8	47.1	45.0
Object-centric-OVD [2]	WS-OVD	Faster R-CNN	36.6	54.0	49.4
VL-PLM	G-OVD	Faster R-CNN	32.3	54.0	48.3
OV-DETR [37]	G-OVD	DeformableDETR[47]	29.4	61.0	52.7
RegionCLIP [42]	C-OVD	CLIP	26.8	54.8	47.5
ViLD [7]	V-OVD	Faster R-CNN	27.6	59.5	51.3
OADP [34]	V-OVD	Faster R-CNN	30.0	53.3	47.2
BARON [35]	V-OVD	Faster R-CNN	34.0	60.4	53.5
BARON†	V-OVD	Faster R-CNN	35.8	58.2	52.3
LBP (ours)	V-OVD	Faster R-CNN	35.9	**60.8**	**54.3**
LBP† (ours)	V-OVD	Faster R-CNN	**37.8**	58.7	53.2

Table 1. Comparison results of LBP and existing SOTA methods on OV-COCO. † indicates model optimization using a batch size of 16, used for mitigating model overfitting towards base classes.

Method	Object detection				Instance segmentation			
	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [7]	16.7	26.5	**34.2**	27.8	16.6	24.6	**30.3**	25.5
DetPro [4]	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
BARON [35]	20.1	28.4	32.2	28.4	19.2	26.8	29.4	26.5
BARON‡	23.2	29.3	32.5	29.5	22.6	27.6	29.8	27.6
LBP (ours)	22.2	28.8	32.4	29.1	22.1	27.0	29.7	27.2
LBP‡ (ours)	**24.1**	**29.5**	32.8	**29.9**	**23.7**	**27.7**	30.1	**28.0**

Table 2. Comparison results of LBP and existing SOTA methods on OV-LVIS. ‡ indicates the model trained using learnable prompt templates, proposed by DetPro [4].

proach lies in V-OVD, but we also present performance results across other settings. Table 1 demonstrates that in this dataset, our LBP approach excels across various benchmark settings, particularly in V-OVD, surpassing existing methods by a notable margin. For instance, compared to BARON†, our method demonstrates a 2.0% improvement in AP_{50}^n and 0.5% in AP_{50}^b in the V-OVD setting, showcasing its efficacy in detecting both base and novel classes during inference. Moreover, it outperforms previous approaches in other settings, affirming its generality and superiority.

Results on OV-LVIS. Compared to OV-COCO, OV-LVIS poses greater challenges due to increased categories and fewer implicit instances for novel classes. The outcomes in Table 2 showcase the exceptional performance of our approach. For the task of object detection, our method enhances BARON's performance by 2.1% on AP_r without learnable prompts and by 0.9% with them. Additionally, our approach consistently boosts the performance of BARON by 2.9% on AP_r in the task of instance segmentation. These comparisons underline the applicability of the proposed method in more complex tasks including object detection and instance segmentation with a larger number of categories, thereby corroborating its effectiveness.

4.3. Ablation Analysis

Impact of individual proposed modules. We conducted an ablation study to assess the effectiveness of each individual module within our method, namely Background Category Prompts (**BCP**), Background Object Discovery

BCP	BOD	IPR	AP_{50}^n	AP_{50}^b	AP_{50}
-	-	-	35.8	58.2	52.3
✓	-	-	34.3	59.0	52.5
-	✓	-	35.2	58.5	52.4
✓	✓	-	35.6	58.7	52.7
✓	-	✓	36.8	59.0	53.2
-	✓	✓	36.9	58.7	53.0
✓	✓	✓	37.8	58.7	53.2

Table 3. Ablation study results in individual proposed modules of LBP on OV-COCO.

Method	AP_{50}^n	AP_{50}^b	AP_{50}
Single embedding	35.8	58.2	52.3
Soft background loss	36.4	58.4	52.6
$\mathcal{L}_{bcp}$	37.3	58.8	53.1
$\mathcal{L}_{rlx}$	36.9	58.4	52.8
LBP (ours)	37.8	58.7	53.2

Table 4. Additional ablation study results of BCP on OV-COCO.

(**BOD**), and Inference Probability Rectification (**IPR**). By incrementally integrating these modules into the full model, we sought to understand their individual impact. Table 3 illustrates the outcomes. We observed that employing only **BCP** notably improves the baseline's performance w.r.t. base classes but results in a decline in performance for novel classes. This observation highlights the inherent conceptual overlap generated between the estimated background underlying classes and novel classes when exclusively relying on **BCP**. However, combining **BCP** with **IPR** demonstrates a 2.5% performance improvement of novel classes, showcasing the effect of **IPR**. Building upon the above variant model, the addition of **BOD** further enhances the model, resulting in a 1.0% mAP50 performance improvement on novel classes. This validates its increased role in identifying objects within the estimated underlying categories from background proposals.

Further analysis of BCP. To further validate **BCP**'s efficacy, we explore various conventional designs for background interpretation in Table 4. We refer "Single embedding" as the design with only a learnable "background" embedding t_{bg}, akin to prior schemes [7, 35], while "Soft background loss" pertains to the soft background loss outlined in DetPro [4]. Subsequently, "$\mathcal{L}_{bcp}$" and "$\mathcal{L}_{rlx}$" indicate our substitution of $\mathcal{L}'_{bcp}$ in Eq. (7) with only $\mathcal{L}_{bcp}$ and $\mathcal{L}_{rlx}$, respectively. The results considerably highlight the superiority of our proposed design for background interpretation. Specifically, our design significantly outperforms "Soft background loss" proposed by DetPro [4], showcasing its effectiveness in preserving class relations. Moreover, the comparison using only $\mathcal{L}_{bcp}$ and $\mathcal{L}_{rlx}$ emphasizes the individual importance of each loss component proposed by **BCP** used for background interpretation.
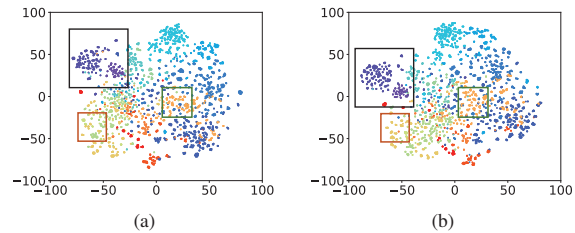

Figure 3. Visualizations of feature distributions for novel category proposals. Note that proposals are selected for those that exhibit significant IoU overlap with the ground truth boxes of novel categories. Different colors denote distinct categories. (a) and (b) showcase the feature distributions generated by BARON [35] and our LBP, respectively. Compared to BARON, our LBP algorithm leads to more compact distributions for the same novel category representations.

Furthermore, we use t-SNE to visualize the feature distribution of novel category proposals, further highlighting the efficacy of our designed schemes in background interpretation. As displayed in Figure 3(a) and Figure 3 (b), compared our approach LBP with BARON [35], the findings illustrate that the proposed LBP approach enables the detector to learn more discriminative features for the proposals associated with novel categories.

5. Conclusions

In this paper, we have introduced LBP, a novel framework addressing the challenges of open-vocabulary object detection. In this approach, learning background prompts is proposed to harness explored implicit background knowledge. This can enhance the capacity of the detecter to recognize both base and novel categories during inference. To achieve this, we have devised three essential modules: Background Category-specific Prompt, Background Object Discovery, and Inference Probability Rectification. These modules collectively empower the detector to discover, represent, and leverage implicit object knowledge explored from background proposals. Our proposed approach has been rigorously evaluated through extensive experiments and thorough ablation studies, confirming its superior performance.

Acknowledgments

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 3:11–12, 2022. 3

[2] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 2, 7

[3] Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[4] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 2, 3, 5, 7, 8

[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 *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 2, 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. In *European Conference on Computer Vision*, pages 266–282. Springer, 2022. 2

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 4, 7, 8

[8] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383, 2023. 3

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

[10] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019. 4

[11] Duojun Huang, Jichang Li, Weikai Chen, Junshi Huang, Zhenhua Chai, and Guanbin Li. Divide and adapt: Active domain adaptation via customized learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7651–7660, 2023. 1

[12] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[16] 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. 2

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

[18] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2505–2514, 2021. 1

[19] Jichang Li, Guanbin Li, Feng Liu, and Yizhou Yu. Neighborhood collective estimation for noisy label identification and correction. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[20] Jichang Li, Guanbin Li, and Yizhou Yu. Adaptive betweenness clustering for semi-supervised domain adaptation. *IEEE Transactions on Image Processing*, 2023.

[21] Jichang Li, Guanbin Li, and Yizhou Yu. Inter-domain mixup for semi-supervised domain adaptation. *Pattern Recognition*, 2023.

[22] Jichang Li, Guanbin Li, Hui Cheng, Zicheng Liao, and Yizhou Yu. Feddiv: Collaborative noise filtering for federated learning with noisy labels. 2024. 1

[23] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

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

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 7

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[27] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 2

[28] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 2

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

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[31] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *arXiv preprint arXiv:2304.04704*, 2023. 3

[32] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[33] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 4

[34] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 2, 7

[35] 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, 4, 5, 7, 8

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

[37] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 7

[38] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[39] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers, 2023. 1

[40] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 2, 4, 5

[41] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2022. 1, 4

[42] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2, 7

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 4

[45] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 7

[46] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2, 7

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 7