From Scene to Object: Enhancing Open-Vocabulary Object Detection via Foreground-Background Context Reasoning

Yanqi Li[1,2], **Jianwei Niu**[1,2,3], **Ningbo Gu**[1,3], **Tao Ren**[4,*]

[1] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China,

[2] Zhongguancun Laboratory, Beijing, China

[3] Hangzhou Innovation Institute of Beihang University, Zhejiang Key Laboratory of Industrial Big Data and Robot Intelligent Systems, Hangzhou, China

[4] State Key Laboratory of Intelligent Game, Institute of Software Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China

{liyanqi_07, niujianwei, guningbo}@buaa.edu.cn, rentao22@iscas.ac.cn

Abstract

Open-Vocabulary Object Detection (OVOD) aims to detect both known and novel categories in complex visual scenes, surpassing the limitations of conventional closed-set detectors. Recent advances in vision-language models (VLMs) like CLIP have enabled zero-shot recognition by aligning visual features with large-scale textual embeddings. However, current OVOD approaches often fall short by overlooking critical contextual and semantic cues necessary for discovering a broader range of novel objects. To address this, we propose **BFDet**, a scene-to-object reasoning framework that leverages the complementary strengths of Large Language Models (LLMs) and VLMs. BFDet introduces a novel scene-to-object reasoning mechanism grounded in foreground-background context interaction. It first uses high-confidence objects to infer the scene-level background. This scene background then guides the discovery of foreground objects by prompting an LLM to generate scene-sensitive novel object candidates. These candidates are subsequently verified through cross-modal alignment and used as high-quality pseudo-labels to enrich detector training. Designed as a plug-and-play module, BFDet integrates seamlessly into existing detection pipelines and consistently improves performance on novel categories across COCO and LVIS benchmarks.

Introduction

Object detection (Ren et al. 2015; Lin et al. 2017; Carion et al. 2020; Xie et al. 2021) a fundamental task in computer vision that has achieved remarkable progress propelled by deep learning (Wu et al. 2024), enabling precise object localization and classification (Papageorgiou and Poggio 2000; He et al. 2017). However, most modern detectors operate under a restrictive *closed-set* assumption, recognizing only categories explicitly defined during training. This assumption stands in stark contrast to the *open-world* nature of real environments, where novel objects are frequently encountered. As a result, these detectors struggle to generalize beyond their pre-defined label space, limiting their real-world applicability (Bansal et al. 2018). In response, *Open-Vocabulary Object Detection* (OVOD) (Kamath et al. 2021; Gu et al. 2021; Zareian et al. 2021; Gao et al. 2022; Li et al. 2022) has emerged as a critical research frontier. OVOD seeks to equip detectors with the ability to recognize novel objects by leveraging knowledge from pre-trained Vision-Language Models (VLMs) like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), which align visual and textual concepts in a shared embedding space to facilitate zero-shot recognition (Zhou et al. 2022; Feng et al. 2022).

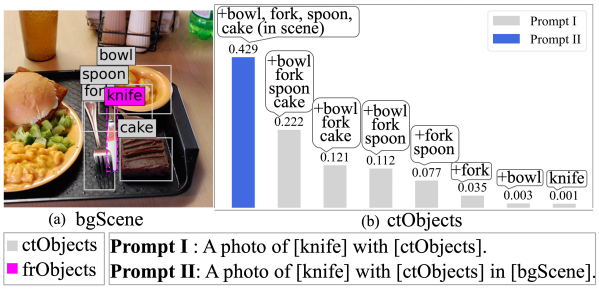


Figure 1: Impact of background scene (bgScene) and contextual objects (ctObjects). Given the few known foreground objects (frObjects) highlighted in *gray boxes*, we analyze the semantic influence of the bgScene on identifying unknown frObjects (*fuchsia box*), and find that the recognition is gradually enhanced by incorporating more accurate ctObjects into bgScene.

While pioneering methods such as (Gu et al. 2021; Wu et al. 2023; Du et al. 2022) improve novel object generalization by distilling VLM representations into detection models, they primarily rely on static visual-textual alignments that treat each object as an isolated visual region matched directly to categories. These methods overlook the surrounding background scene and contextual objects, which often provide essential semantic cues for clarifying object identity. As a result, these methods struggle to capture the implicit knowledge embedded in background regions, includ-

*Corresponding author: Tao Ren.

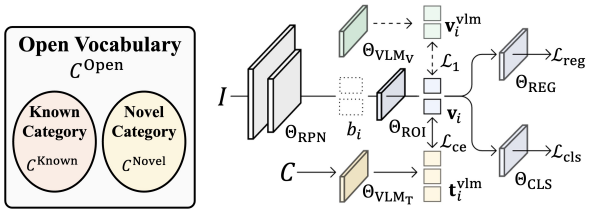
ing contextual object relationships and global scene semantics, leading to critical limitations in background understanding. Building on this, recent advances (Bangalath et al. 2022; Feng et al. 2022; Gao et al. 2022; Li et al. 2024) attempt to mine implicit novel objects through pseudo-labeling proposals across the scene. However, they mainly focus on foreground content overlooking the broader semantic context essential for scene understanding. As a result, contextual objects and scene-level knowledge remain underused due to the absence of explicit background reasoning.

As shown in Figure 1(a), the object "*knife*" (fuchsia box) is an unknown object (**foreground object, frObject**) in OV-COCO (Lin et al. 2014). CLIP fails to recognize this object alone in the region semantics (**background scene, bgScene**), as semantic ambiguity arises in its visual features when detached from contextual surroundings (**contextual objects, ctObjects**). In practice, objects rarely appear in isolation. Comprehensive background semantics, including bgScene and ctObjects, serves as global priors that guide recognition on novel categories and disambiguate visually similar categories. Therefore, we conducted an experiment (Figure 1(b)) to investigate the effect of bgScene and ctObjects on object recognition accuracy in CLIP. Results show that the progressive improvement in CLIP's recognition of the frObjects is largely attributed to the increasing availability of accurate bgScene and ctObjects. This insight highlights the critical role of background in OVOD: If high-confidence *bgScene* and *ctObjects* are available, it critically enhances CLIP's recognition accuracy for novel *frObjects*, thereby boosting OVOD. Accordingly, this motivates a key question: *How can we identify and utilize high-confidence background context?*

To address this, we explore the interaction between Large Language Models (LLMs) and VLMs to construct reliable ctObject-enhanced background knowledge, where the knowledge inferred by LLMs is cross-modally validated through the semantic alignment of VLMs. LLMs possess rich high-level semantic knowledge and textual reasoning capabilities grounded in world knowledge, enabling them to infer plausible frObjects based on the given bgScene and ctObjects. In contrast, VLMs excel at aligning visual and textual semantics. Yet, given that only a partial set of known ctObjects is available in OVOD, two fundamental challenges arise:

- **Contextual Background Reasoning:** How can we infer and enrich bgScene semantics from initial limited known ctObjects?
- **Cross-Modal Object Grounding:** How to discover then ground novel objects guided by enriched background?

To tackle these challenges, we develop **BFDet**, a **B**ackground-to-**F**oreground reasoning framework that enhances open-vocabulary object **Det**ection by incorporating high-level semantic reasoning from LLMs and visual grounding from VLMs. Specifically, BFDet first infers the underlying bgScene from high-confidence ctObjects using LLMs, and then incorporates reliable ctObjects into bgScene to guide the discovery of frObjects. Candidate novel frObjects are verified and grounded through VLM-based visual-text alignment, and the resulting object-category pairs are used as pseudo-labels to supervise detector training, enhancing novel object recognition. BFDet is architecture-agnostic and can be flexibly integrated into any VLM-based detection framework. Experiments on COCO and LVIS benchmarks show that BFDet leads to a +3.1 AP improvement in detecting novel categories while maintaining strong performance on known ones. Our contributions are as follows:



(a) Category relation. (b) VLM-based OVOD architecture.

Figure 2: (a) Category relationships in the problem definition. (b) Typical architecture of VLM-based OVOD.

- We propose a novel scene-to-object reasoning framework that leverages foreground-background contextual information to enhance novel category detection in OVOD.
- We design a knowledge discovery mechanism that leverages limited known objects to uncover rich background scene cues for further novel object reasoning.
- We introduce a cross-modal alignment method that validates inferred background scene and grounds novel objects, producing reliable pseudo-labels to improve the detector's generalization to new categories.
- We demonstrate state-of-the-art performance on both OV-COCO and OV-LVIS benchmarks through extensive experiments and ablation studies that confirm the effectiveness of our proposed method.

Preliminary

Problem Definition

OVOD aims to train a detector using predefined *known categories* $\mathcal{C}^{\text{Known}}$, while enabling it to generalize to *novel categories* $\mathcal{C}^{\text{Novel}}$ at inference. Notably, the known and novel category sets are mutually exclusive, i.e., $\mathcal{C}^{\text{Known}} \cap \mathcal{C}^{\text{Novel}} = \emptyset$.

The training dataset $\mathcal{D}$ consists of image-annotation pairs (I, A), where each annotation $\{(b_i, c_i)\} \in A$ denotes a bounding box b_i and a corresponding label $c_i \in \mathcal{C}^{\text{Known}}$. In addition, the dataset may include open-vocabulary textual descriptions $\mathcal{C}^{\text{Open}}$ to provide richer semantic supervision. During inference, the detector is evaluated on its ability to localize objects and assign category labels from the unified label space $\mathcal{C} = \mathcal{C}^{\text{Known}} \cup \mathcal{C}^{\text{Novel}}$.

VLM-based OVOD

OVOD methods (Gu et al. 2021; Ma et al. 2022) enhance detectors by aligning region features with visual and textual embeddings in a shared semantic space, leveraging pre-trained VLMs for zero-shot recognition (Figure 2(b)).

Given an image I, proposals b_i from a Region Proposal Network (Θ_{RPN}) are used to extract ROI features $\mathbf{v}_i = \Theta_{\text{ROI}}(I, b_i)$. These features are then aligned with VLM-derived visual embeddings $\mathbf{v}_i^{\text{vlm}} = \Theta_{\text{VLM}_\text{V}}(b_i)$ and textual embeddings $\mathbf{t}_i^{\text{vlm}} = \Theta_{\text{VLM}_\text{T}}(c_i)$.

Training: The total loss function is:

$$\mathcal{L}_{\text{OVOD}} = \mathcal{L}_1 + \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}. \tag{1}$$

Each component of the loss is defined as follows:

1. **Classification Loss** ($\mathcal{L}_{\text{cls}}$): Encourages accurate category prediction for region features using the standard cross-entropy loss (ℓ_{ce}) (Zhang and Sabuncu 2018). $\mathcal{L}_{\text{cls}} = \sum_i (\ell_{\text{ce}}(\Theta_{\text{CLS}}(v_i), c_i))$.
2. **Regression Loss** ($\mathcal{L}_{\text{reg}}$): Refines the localization of predicted bounding boxes with respect to the ground-truth bounding box b_i^* using the L1 loss (ℓ_{L_1}). $\mathcal{L}_{\text{reg}} = \sum_i (\ell_{L_1}(\Theta_{\text{REG}}(\mathbf{v}_i), b_i^*))$.
3. **Visual Distillation Loss** ($\mathcal{L}_1$): Aligns the detector's visual features with the visual embeddings from the VLM via L1 norm $\|\cdot\|_1$. $\mathcal{L}_1 = \sum_i \|\mathbf{v}_i - \mathbf{v}_i^{\text{vlm}}\|_1$.
4. **Textual Distillation Loss** ($\mathcal{L}_{\text{ce}}$): Acts as a contrastive loss that aligns detector features with VLM's textual embeddings. $\mathcal{L}_{\text{ce}} = \sum_i \ell_{\text{ce}}(\text{sim}(\mathbf{v}_i, \mathbf{t}^{\text{vlm}}), c_i)$, where $\text{sim}(\cdot,\cdot)$ denotes a similarity function.

Inference: The detector predicts regions and categories by:

$$\{\mathbf{v}_i\}_{i=1}^M = \Theta_{\text{ROI}}(I, b_i), \tag{2}$$

$$\{b_i, c_i\}_{i=1}^M = \{\Theta_{\text{REG}}(\mathbf{v}_i), \Theta_{\text{CLS}}(\mathbf{v}_i)\}_{i=1}^M, \tag{3}$$

where $c_i \in \{\mathcal{C}^{\text{Known}} \cup \mathcal{C}^{\text{Novel}}\}$.

Method

Overview

As illustrated in Figure 3, BFDet enhances novel object detection via foreground-background contextual reasoning and visual-textual alignment. **Scene-Aware Background Reasoning**: For each high-confidence proposal b_i^{h} and its surrounding region r_{b_i}, BFDet constructs a background prompt using high-confidence categories c^{h} to infer the bgScene c^{bg}. **Foreground Objects Alignment**: It then predicts frObject categories c^{fr} within r_{b_i} by constructing an object prompt combining c^{h} and c^{bg}, and aligning it with visual features. **Contextual Knowledge Injection**: The inferred c^{fr} serve as pseudo-labels for low-confidence proposals b^{l}, providing external supervision to enhance novel categories detection.

Scene-Level: Scene-Aware Background Reasoning

BFDet extracts high-confidence ctObjects b^{h} within the expanded region r_{b_i} to prompt candidate bgScenes and selects the probable background c^{bg} by aligning textual semantics with visual features.

CtObjects Extraction. To obtain bgScene, BFDet expands the object proposal b_i by a factor $\gamma = 2$ to define r_{b_i}. Formally, for $b_i = [x_1, y_1, x_2, y_2]$:

$$r_{b_i} = \left[(c_x - \frac{\gamma w}{2}, c_y - \frac{\gamma h}{2}), (c_x + \frac{\gamma w}{2},\ c_y + \frac{\gamma h}{2})\right], \tag{4}$$

where $w = x_2 - x_1$, $h = y_2 - y_1$. (c_x, c_y) is box center.

Within r_{b_i}, objects are split into ctObjects $\mathcal{B}^{\text{h}}$ and low-confidence objects $\mathcal{B}^{\text{l}}$ using a threshold θ. An object belongs to $\mathcal{B}^{\text{h}}$ if its maximum probability $p_{b,c} > \theta$, where we set $\theta = 0.6$. This yields ctObjects:

$$\mathcal{B}^{\text{h}} = \left\{ b \subseteq r_{b_i} \mid \max_{c \in \mathcal{C}^{\text{Known}}} p_{b,c} > \theta \right\}. \tag{5}$$

For each $(b^{\text{h}}, c^{\text{h}}) \in \mathcal{B}^{\text{h}}$, the category c^{h} is determined by $\arg\max_c(p_{b^{\text{h}},c})$, where $p_{b,c}$ is the softmax probability:

$$p_{b,c} = \frac{\exp(\mathbf{w}_c^\top f_b)}{\sum_{c' \in \mathcal{C}^{\text{Known}}} \exp(\mathbf{w}_{c'}^\top f_b)}. \tag{6}$$

Here, $f_b \in \mathbb{R}^d$ is the embedding of b, and $\mathbf{w}_c \in \mathbb{R}^d$ is the weight vector for category c. These ctObject pairs are used for subsequent reasoning.

BgScene Reasoning. We leverage ctObject categories $\mathcal{C}^{\text{h}}$ to estimate bgScene c^{bg} for region r_{b_i}. Assuming objects in the region share a common background, we construct the following *background prompt* to query the LLM for plausible background $\mathcal{C}^{\text{bg}}$, where $|\mathcal{C}^{\text{bg}}| = N^{\text{bg}}$.

> *Identify* $[N^{\text{bg}}]$ *typical background scenes where* $[\mathcal{C}^{\text{h}}]$ *usually appear together.*

We utilize VLMs to achieve visual-text alignment of the background. For a background $c_i^{\text{bg}} \in \mathcal{C}^{\text{bg}}$, we compute:

- Visual features: $\mathbf{v}^{\text{bg}} = \Theta_{\text{VLM}_\text{V}}(r'_{b_i})$, where r'_{b_i} is the region r_{b_i} excluding objects $\mathcal{B}^{\text{l}}$.
- Textual features: $\mathbf{t}_i^{\text{bg}} = \Theta_{\text{VLM}_\text{T}}(text_i^{\text{bg}})$, where $text_i^{\text{bg}} =$ `A photo of` $\mathcal{C}^{\text{h}}$ `in` c_i^{bg}.

We then compute cosine similarity by Eq. 7 and convert it into a confidence score using sigmoid function by Eq. 8:

$$\cos(\mathbf{v}^{\text{bg}}, \mathbf{t}_i^{\text{bg}}) = \frac{(\mathbf{v}^{\text{bg}})^\top \mathbf{t}_i^{\text{bg}}}{\|\mathbf{v}^{\text{bg}}\| \|\mathbf{t}_i^{\text{bg}}\|}, \tag{7}$$

$$p(c_i^{\text{bg}}) = \sigma\big(\cos(\mathbf{v}^{\text{bg}}, \mathbf{t}_i^{\text{bg}})\big) = \frac{1}{1 + \exp\big(-\cos(\mathbf{v}^{\text{bg}}, \mathbf{t}_i^{\text{bg}})\big)}. \tag{8}$$

A background c_i^{bg} is selected as reliable if its confidence score $p(c_i^{\text{bg}}) > \alpha$, where $\alpha = 0.6$.

Object-Level: Foreground Objects Alignment

Foreground Category Reasoning. Given ctObjects $(b^{\text{h}}, c^{\text{h}})$ and background c^{bg}, we aim to infer the categories $\mathcal{C}^{\text{l}}$ of objects $\mathcal{B}^{\text{l}}$ in r_{b_i}. To this end, we leverage c^{bg} to predict frObjects that are likely to appear in the scene by querying LLMs with a *objects prompt*:

> *In* $[c^{\text{bg}}]$, *list* $[N^{\text{fr}}]$ *common objects that usually co-occur nearby* $[c^{\text{h}}]$.

We query the LLM N^{q} times to generate candidate frObjects $\mathcal{C}^{\text{fr}}$, where $|\mathcal{C}^{\text{fr}}| = N^{\text{fr}}$. For each set of candidate objects $c_i^{\text{fr}} \in \mathcal{C}^{\text{fr}}$, we construct a descriptive context ($text_i^{\text{fr}}$) for subsequent alignment with visual features as follows,

`There are` c^{h} `and` c_i^{fr} `in` c^{bg}.

Next, we align the contextual descriptions with low-confidence proposals to assign accurate semantic labels.

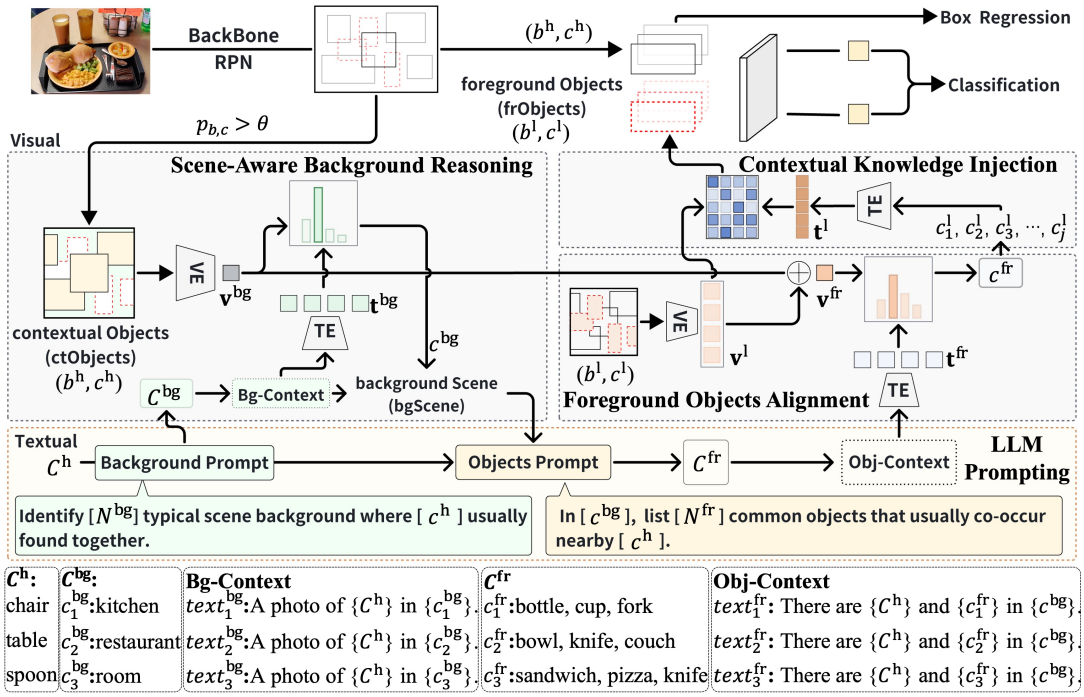


Figure 3: The BFDetframework enhances OVOD via a two-stage foreground-background reasoning process. It first uses high-confidence ctObjects $\mathcal{B}^{h}$ to prompt an LLM for potential bgScenes C^{bg}, which are then validated via VLM-based alignment. The verified bgScene c^{bg} guides a second LLM prompt to generate likely frObject categories C^{fr} as object-level context. These are matched to low-confidence proposals $\mathcal{B}^{l}$ for reliable pseudo-label assignment. The resulting external knowledge is injected into the training process, enhancing novel object recognition.

Object and Category Alignment. To identify plausible category set for the low-confidence proposals in region r_{b_i}, we leverage a VLM to align each candidate c_i^{fr} with the region's visual features using cross-modal alignment.

- Visual Embedding: $\mathbf{v}^{fr} = \Theta_{\text{VLM}_V}(r_{b_i})$, capturing bgScene and low-confidence ctObjects.
- Text Embedding: $\mathbf{t}_i^{fr} = \Theta_{\text{VLM}_T}(text_i^{fr})$, encoding the contextual description.

The probability is measured using cosine similarity (Eq. 9):

$$p(c_i^{fr}) = \sigma\big(\cos(\mathbf{v}^{fr}, \mathbf{t}_i^{fr})\big) = \frac{1}{1+\exp\big(-\cos(\mathbf{v}^{fr}, \mathbf{t}_i^{fr})\big)}. \quad (9)$$

The category set c_i^{fr} is selected as a pseudo-label pool for low-confidence proposals $\mathcal{B}^{l}$ if its score satisfies $p(c_i^{fr}) > \beta$, where $\beta = 0.5$.

Contextual Knowledge Injection

Object-Category Matching. Given the set of validated frObject categories c^{fr}, the goal is to assign the single best category to each individual low-confidence object $b_i^{l} \in \mathcal{B}^{l}$ to identify reliable pseudo-label pairs. This is treated as an object-level matching problem.

For each object proposal b_i^{l}, we first extract its visual features $\mathbf{v}_i^{l} = \Theta_{\text{VLM}_V}(b_i^{l})$ and then compare these features against the textual embedding $\mathbf{t}_j^{l} = \Theta_{\text{VLM}_T}(c_j^{l})$ of every candidate category $c_j^{l} \in c^{fr}$. This visual-textual alignment is measured using cosine similarity:

$$s_{ij} = \cos(\mathbf{v}_i^{l}, \mathbf{t}_j^{l}) = \frac{(\mathbf{v}_i^{l})^{\top}\mathbf{t}_j^{l}}{\|\mathbf{v}_i^{l}\| \cdot \|\mathbf{t}_j^{l}\|}. \quad (10)$$

To determine the most likely label for the object b_i^{l}, we convert these similarity scores into a probability distribution over all candidate categories using the softmax function:

$$P(c_j^{l} \mid b_i^{l}) = \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})}, \quad \text{for all } c_k^{l} \in c^{fr}. \quad (11)$$

The category with the top score is selected as the pseudo-label only if its confidence exceeds a threshold ω:

$$c_i^{l} = \arg\max_{c_j^{l} \in c^{fr}} P(c_j^{l} \mid b_i^{l}), \quad \text{if } \max_j P(c_j^{l} \mid b_i^{l}) > \omega. \quad (12)$$

Here, we set $\omega = 0.6$. This ensures that high-confidence frObjects are retained, forming reliable pseudo-label pairs $\{(b_i^{l}, c_i^{l})\}$ for training.

Knowledge Supervision. With a set of reliable pseudo-labels $\{(b_i^{l}, c_i^{l})\}$, we can inject this contextual knowledge into the object detection framework. We introduce a classification loss $\mathcal{L}_{\text{BFDet}}$ that supervises the model using these newly identified pairs:

$$\mathcal{L}_{\text{BFDet}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(p_{\text{cls}}(c_i^{l} \mid \mathbf{v}_i^{l})\right), \quad (13)$$

where N is the number of identified pseudo-label pairs. The loss promotes alignment between the visual embedding $\mathbf{v}_i^{\mathrm{l}}$ of each frObject and the textual embedding of its assigned pseudo-label c_i^{l}. The probability p_{cls} is computed via a temperature-scaled softmax, which sharpens the output distribution to amplify the learning signal:

$$p_{\mathrm{cls}}(c \mid \mathbf{v}_i^{\mathrm{l}}) = \frac{\exp(\cos(\mathbf{v}_i^{\mathrm{l}}, \mathbf{t}_c)/\tau)}{\sum_{k=1}^{|\mathcal{C}|} \exp(\cos(\mathbf{v}_i^{\mathrm{l}}, \mathbf{t}_k)/\tau)}. \quad (14)$$

Here, $\mathbf{t}_c$ is the textual embedding for a category c, and τ is the temperature parameter. By optimizing $\mathcal{L}_{\mathrm{BFDet}}$, BFDet learns the semantic features of novel objects, bridging the gap between known and novel categories.

The final training objective for the enhanced detector is the sum of the original loss and our classification loss:

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{OVOD}} + \mathcal{L}_{\mathrm{BFDet}}. \quad (15)$$

Experiments

We present the experiments in this section. More experimental details and results are provided in Appendix B.

Datasets

We evaluate BFDet on two widely adopted OVOD benchmarks: COCO (Lin et al. 2014) and LVIS (Gupta, Dollar, and Girshick 2019).

OV-COCO (COCO Benchmark). (1) *Category Split*: 48 known and 17 novel categories (following OVR-CNN (Zareian et al. 2021)). (2) *Train Set*: 107,761 images annotated with known categories (Zareian et al. 2021). (3) *Validation Set*: 4,836 images with both known and novel category annotations (Zareian et al. 2021). (4) *Metric*: Average Precision (AP) at IoU=0.5 (Zareian et al. 2021; Gu et al. 2021).

OV-LVIS (LVIS Benchmark): (1) *Category Split*: 866 known (common + frequent) and 337 novel (rare) categories, following ViLD (Gu et al. 2021). (2) *Metric*: mask AP (mAP) following official LVIS protocol (Lin et al. 2022).

Implementation Details

On OV-COCO, we use Faster R-CNN (Ren et al. 2015) with a ResNet-50-C4 (He et al. 2016) backbone and SGD optimizer, a batch size of 8, and warm-up to 2e-3 over 1k iterations, followed by 10× learning rate decay at 6k and 8k iterations. On OV-LVIS, we adopt CenterNet2 (Zhou, Koltun, and Krähenbühl 2021) with a ResNet-50 backbone and Adam optimizer, using a batch size of 8 and warming up to 2e-4 over 1k iterations. In all settings, CLIP's text encoder is used to generate embeddings for both known and novel categories inferred via foreground-background reasoning. For method-specific parameter settings, BFDet prompts $N^{\mathrm{bg}} = 5$ bgScene candidates and $N^{\mathrm{fr}} = 5$ frObjects, with $N^{\mathrm{q}} = 3$ queries for foreground reasoning. We adapt CLIP as VLM encoder and query to Qwen-3 (Yang et al. 2025).

Integration with Released Methods. For methods with publicly available code, we apply BFDet to OV-COCO using five SOTA VLM-based detectors: ViLD (Gu et al. 2021), Detic (Zhou et al. 2022), RegionCLIP (Zhong et al. 2022), BARON (Wu et al. 2023), and VLDet (Lin et al. 2022). On OV-LVIS, BFDet is integrated into four detectors across two backbones: ViLD, RegionCLIP, and VLDet with ResNet-50 (RN50); Detic and VLDet with Swin-B.

Method	AP^{Novel}	AP^{Known}	AP^{All}
Base	1.3	52.8	39.3
OC-OVD [NeurIPS, 2022]	36.6	54.0	49.4
OADP [ICCV, 2023]	36.4	53.0	48.6
CoDet [NeurIPS, 2023]	30.6	52.3	46.6
LBP [CVPR, 2024]	<u>37.8</u>	58.7	<u>53.2</u>
CAKE [AAAI, 2025]	**39.1**	58.1	53.1
ViLD [ICLR, 2022]	27.6	<u>59.5</u>	51.3
ViLD-BFDet *(Ours)*	29.9(2.3)	**62.1**(2.6)	**53.7**(2.4)
Detic [ECCV, 2022]	27.8	51.1	45.0
Detic-BFDet *(Ours)*	30.3(2.5)	54.4(3.3)	47.9(2.9)
RegionCLIP [CVPR, 2022]	26.8	54.8	47.5
RegionCLIP-BFDet *(Ours)*	29.5(2.7)	58.0(3.2)	50.4(2.9)
VLDet [ICLR, 2023]	32.0	50.6	45.8
VLDet-BFDet *(Ours)*	35.1(3.1)	54.1(3.5)	49.0(3.2)
BARON [CVPR, 2023]	35.1	54.8	49.1
BARON-BFDet *(Ours)*	<u>37.8</u>(2.7)	58.2(3.4)	52.0(2.9)

Table 1: On OV-COCO, BFDet achieves consistent gains over SOTA in both novel (AP^{Novel}) and known (AP^{Known}) categories. The Base method refers to Faster R-CNN trained with CLIP embeddings on COCO known categories.

Comparison with Other Methods. We further compare BFDet with several recent OVOD approaches, such as DetPro (Du et al. 2022), OC-OVD (Rasheed et al. 2022), CoDet (Ma et al. 2023), OADP (Wang et al. 2023), LBP (Li et al. 2024), and CAKE (Ma et al. 2025), using reported results on OV-COCO and OV-LVIS. Note that DetPro does not report results on OV-COCO.

Comparison with SOTA

Experimental Results on OV-COCO. As shown in Table 1, BFDet consistently enhances SOTA VLM-based detectors (with available official codebases) across both novel and known categories. VLDet-BFDet achieves 35.1 AP^{Novel} (+3.1 over VLDet) and 54.1 AP^{Known} (+3.5), representing the largest overall gains. ViLD-BFDet achieves 62.1 AP^{Known} and 53.7 AP^{All} yielding the best overall performance. Meanwhile, BARON-BFDet achieves 37.8 AP^{Novel} that is only marginally behind the SOTA (39.1 AP^{Novel}). These improvements are attributed to BFDet's ability to incorporate background context knowledge as external guidance, effectively bridging the gap between known and novel categories. This enables more accurate object-category pairs, covering not only known but also novel categories. Importantly, BFDet operates as a plug-and-play module, requiring no changes to the underlying detection architecture, highlighting its versatility and scalability.

Experimental Results on OV-LVIS. On the challenging OV-LVIS benchmark (Table 2), generalizing to novel objects remains a significant hurdle, evidenced by the base-

Method	Backbone	mAP^{mask}_{Novel}	mAP^{mask}_{Common}	$mAP^{mask}_{Frequent}$	mAP^{mask}_{All}
Base	RN50	16.3	31.0	35.4	30.0
DetPro [CVPR, 2022]	RN50	20.8	27.8	32.4	28.4
OC-OVD [NeurIPS, 2022]	RN50	21.1	25.0	29.1	25.9
OADP [ICCV, 2023]	RN50	23.3	29.7	34.3	30.4
CoDet [NeurIPS, 2023]	RN50	23.4	30.0	34.6	30.7
LBP [CVPR, 2024]	Swin-B	24.1	29.5	32.8	29.9
CAKE [AAAI, 2025]	RN50	25.0	34.8	38.4	34.9
ViLD [ICLR, 2022]	RN50	16.6	24.6	30.3	25.5
ViLD-BFDet *(Ours)*	RN50	19.1 (2.5)	26.2 (1.6)	31.9 (1.6)	27.5 (2.0)
RegionCLIP [CVPR, 2022]	RN50	17.1	27.4	34.0	28.2
RegionCLIP-BFDet *(Ours)*	RN50	19.5 (2.4)	28.7 (1.3)	35.8 (1.8)	29.8 (1.6)
VLDet [ICLR, 2023]	RN50	21.7	29.8	34.3	30.1
VLDet-BFDet *(Ours)*	RN50	23.6 (1.9)	31.6 (1.8)	36.1 (1.8)	31.9 (1.8)
BARON [CVPR, 2023]	RN50	19.2	26.8	29.4	26.5
BARON-BFDet *(Ours)*	RN50	20.7 (1.5)	28.7 (1.9)	30.8 (1.4)	28.3 (1.8)
Detic [ECCV, 2022]	Swin-B	23.9	40.2	42.8	38.4
Detic-BFDet *(Ours)*	Swin-B	25.4 (1.5)	**41.6** (1.4)	**44.7** (1.9)	**40.1** (1.7)
VLDet [ICLR, 2023]	Swin-B	26.3	39.4	41.9	38.1
VLDet-BFDet *(Ours)*	Swin-B	**28.1** (1.8)	41.0 (1.6)	43.2 (1.3)	39.8 (1.7)

Table 2: Performance on OV-LVIS benchmark, BFDet consistently improves mask AP for novel (mAP^{mask}_{Novel}) and known (split into Common and Frequent subsets) categories across SOTA methods and backbones (RN50/Swin-B). The Base method refers to training only on known categories.

line's low 16.3 mAP^{mask}_{Novel}. Our method, BFDet, is designed as a plug-and-play module to address this gap. When integrated with strong state-of-the-art (SOTA) detectors, it yields substantial improvements across diverse backbones. For example, it boosts ViLD (RN50) by +2.5 and, most notably, elevates VLDet (Swin-B) to a new SOTA performance of 28.1 mAP^{mask}_{Novel} (+1.8). Crucially, it demonstrates even greater dominance on known categories. When paired with Detic, BFDet establishes new best-in-class results across the board, achieving 41.6 mAP^{mask}_{Common} (+1.4), 44.7 in $mAP^{mask}_{Frequent}$ (+1.9) and 40.1 in mAP^{mask}_{All} (+1.7). This comprehensive improvement validates that leveraging background context is a highly effective strategy for bridging the gap between known and novel categories without requiring architectural modifications.

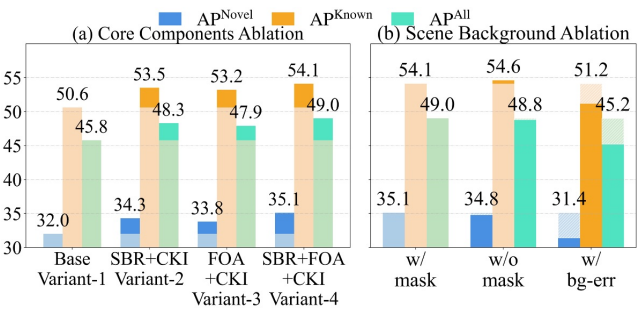


Figure 4: Ablation studies of BFDet performed on OV-COCO with VLDet as the base model, evaluated with AP^{Novel}, AP^{Known} and AP^{All}.

Ablation Studies

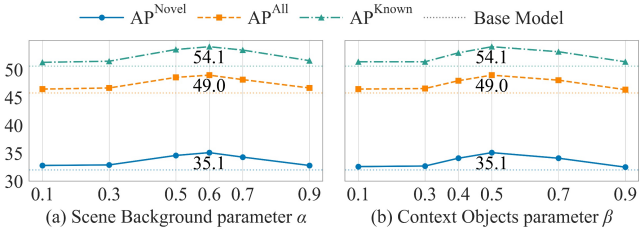
Core Components. We conducted an ablation study to evaluate the contribution of Scene-Aware Background Reasoning (**SBR**), Foreground Objects Alignment (**FOA**), and Contextual Knowledge Injection (**CKI**), respectively. These components are incrementally integrated into the framework to analyze their individual effects. Notably, SBR and FOA rely on CKI where disabling CKI renders them ineffective. Conversely, without SBR and FOA to extract ctObjects, CKI alone fails to provide meaningful improvements. Therefore, we did not include comparisons for these combinations. As shown in Figure 4(a), Variant 1 corresponds to the base detector VLDet without any added module. Variants 2 vs. 3 indicate that SBR contributes more significantly to performance gains than FOA. Comparing Variants 2 and 3 to Variants 1 and 4 indirectly underscores the importance of bgScene understanding and frObjects reasoning for enhancing novel object detection.

BgScene Influence. As illustrated in Figure 4(b), we compare two variants, "*w/ mask*" vs. "*w/o mask*", for identifying the background representation $\mathbf{v}^{bg}$ in Eq. (8) to select a reliable background label c^{bg}. Our method, "*w/ mask*", embeds the region after masking out low-confidence objects $\mathcal{B}^{l}$. As a baseline, "*w/o mask*" directly embeds the entire unmasked region as $\mathbf{v}^{bg} = \Theta_{\text{VLM}_\text{V}}(r_{b_i})$. We also compare an ablation variant "*w/ bg-err*" with "*w/ mask*", which introduces an incorrect background into the *Objects Prompt* and $text^{fr}_i$ to assess its impact on object reasoning. Variant "*w/ mask*" achieves the best results, with an AP^{Novel} of 35.1 and an overall AP^{All} of 49.0. Compared to this, "*w/o*

(a) Identified novel categories within COCO label space.

(b) Identified novel categories beyond the COCO label space.

Figure 5: Visualization of pseudo-labels. *Gray* boxes denote known categories in the COCO annotations, *yellow* boxes highlight novel categories that exist in COCO labels, and *blue* boxes mark novel categories that go beyond the COCO label space.



(a) Scene Background parameter α (b) Context Objects parameter β

Figure 6: Hyperparameter analysis of BFDet performed on OV-COCO with VLDet as base model.

mask" shows a slight gain on known categories (AP^{Known}: +0.5 AP) but a minor drop on novel categories (AP^{Novel}: -0.3 AP), suggesting that unfiltered low-confidence objects may introduce semantic noise, slightly hindering novel object recognition which relies more on clean context. The most notable degradation appears in "*w/ bg-err*", where using incorrect background leads to sharp performance drops: AP^{Novel} drops by -3.7 and AP^{All} by -3.8. This shows that accurate bgScene is essential for effective object reasoning, as misleading context negatively affects the recognition of both novel and known objects.

Visualization of Pseudo-Labels

As illustrated in Figure 5, we visualize the pseudo-labels identified by our background reasoning mechanism. BFDet successfully identifies labels for novel objects that span the entire COCO vocabulary (both known and novel categories) and even for objects beyond it. For example, in Figure 5(a), BFDet successfully infers the novel category "*bus*" by leveraging surrounding ctObjects such as "*car*" and "*person*", along with the bgScene "*road*". It also identifies the known but unlabeled category "*bench*" (highlighted in the yellow box) based on ctObjects like "*bed*" and "*bench*", as well as the bgScene "*bedroom*". Notably, all these categories are part of the original COCO annotations. In contrast, Figure 5(b) illustrates BFDet's ability to infer novel categories beyond the COCO annotations. For instance, it identifies "*plate*" by utilizing contextual cues such as "*cake*" and "*spoon*" within "*table*" bgScene, and infers "*cabinet*" through surrounding objects like "*oven*", "*microwave*", and "*refrigerator*" in a "*kitchen*". This suggests that BFDet can effectively leverage contextual reasoning to discover and label a diverse set of objects, expanding the detector's knowledge far beyond its initial training data. Therefore, these accurate pseudo-labels can enhance the novel category detection.

Hyperparameter Analysis on α and β Thresholds

We investigate the impact of two key hyperparameters, α and β, which collaboratively guide our bgScene and ctObject reasoning process, as shown in Figure 6. The parameter α is responsible for selecting the most relevant bgScene, with our model achieving peak performance at $\alpha = 0.6$. An overly high α is too restrictive, hindering the selection of a suitable scene, while an excessively low value risks choosing an irrelevant scene that misleads subsequent reasoning. Once a scene is identified, the parameter β governs the trade-off between the quality and quantity of the context objects selected from it, with optimal performance observed at $\beta = 0.5$. A high β ensures high-quality objects but limits their quantity, thereby constraining the available contextual information. Conversely, a low β introduces noisy or irrelevant objects that degrade performance. Thus, robust reasoning requires a careful co-tuning of both parameters: α to identify the correct context and β to populate that context with relevant, high-quality objects.

Conclusion

We introduced BFDet, a novel scene-to-object reasoning framework that enhances OVOD by synergizing LLMs and VLMs. Our method leverages foreground-background context, using high-confidence detections to infer a background scene that in turn guides an LLM to hypothesize novel objects. These hypotheses are validated via cross-modal alignment to identify high-quality pseudo-labels. As a plug-and-play module, BFDet consistently boosts the performance of existing detectors on novel categories across benchmarks like COCO and LVIS, demonstrating the power of contextual reasoning for open-world recognition.

Limitations. The performance hinges on the capabilities of the underlying LLMs and VLMs. Its multi-step design introduces computational overhead and potential error propagation. Future work will focus on developing end-to-end training for a more unified model, incorporating richer contextual representations that capture scene and object relationships, and extending to more complex domains to leverage comprehensive context for robust novel object discovery.

Acknowledgments

This work was supported by National Key R&D Program of China (2023YFB4503700), National Natural Science Foundation of China under Grant No. 62372027, U23B2025.

References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 384–400.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14084–14093.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, 701–717. Springer.

Gao, M.; Xing, C.; Niebles, J. C.; Li, J.; Xu, R.; Liu, W.; and Xiong, C. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, 266–282. Springer.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1780–1790.

Li, J.; Zhang, J.; Li, J.; Li, G.; Liu, S.; Lin, L.; and Li, G. 2024. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16678–16687.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2022. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Ma, C.; Jiang, Y.; Wen, X.; Yuan, Z.; and Qi, X. 2023. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36: 71078–71094.

Ma, S.; Qian, D.; Ye, K.; and Zhang, S. 2025. Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5982–5990.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14074–14083.

Papageorgiou, C.; and Poggio, T. 2000. A trainable system for object detection. *International journal of computer vision*, 38: 15–33.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rasheed, H.; Maaz, M.; Khattak, M. U.; Khan, S.; and Khan, F. S. 2022. Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection. In *36th Conference on Neural Information Processing Systems (NIPS)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Wang, J.; Zhang, H.; Hong, H.; Jin, X.; He, Y.; Xue, H.; and Zhao, Z. 2023. Open-vocabulary object detection with an

open corpus. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6759–6769.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15254–15264.

Wu, X.; Liu, X.; Niu, J.; Wang, H.; Tang, S.; Zhu, G.; and Su, H. 2024. Decoupling General and Personalized Knowledge in Federated Learning via Additive and Low-rank Decomposition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 7172–7181. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8392–8401.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14393–14402.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, 350–368. Springer.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2021. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*.