

Benefit From Seen: Enhancing Open-Vocabulary Object Detection by Bridging Visual and Textual Co-Occurrence Knowledge

Yanqi Li^{1,2} Jianwei Niu^{1,2,3} Tao Ren^{4,*}

¹State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China ²Zhongguancun Laboratory, Beijing, China

³Hangzhou Innovation Institute, Beihang University, Hangzhou, China

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

{liyanqi-07, niujianwei}@buaa.edu.cn, rentao22@iscas.ac.cn

Abstract

Open-Vocabulary Object Detection (OVOD) aims to localize and recognize objects from both known and novel categories. However, existing methods rely heavily on internal knowledge from Vision-Language Models (VLMs), restricting their generalization to unseen categories due to limited contextual understanding. To address this, we propose CODet, a plug-and-play framework that enhances OVOD by integrating object co-occurrence — a form of external contextual knowledge pervasive in real-world scenes. Specifically, CODet extracts visual co-occurrence patterns from images, aligns them with textual dependencies validated by Large Language Models (LLMs), and injects contextual co-occurrence pseudo-labels as external knowledge to guide detection. Without architectural changes, CODet consistently improves five state-of-the-art VLM-based detectors across two benchmarks, achieving notable gains (up to +2.3 AP on novel categories). Analyses further confirm its ability to encode meaningful contextual guidance, advancing open-world perception by bridging visual and textual co-occurrence knowledge.

1. Introduction

Object detection [28, 32], a core task in computer vision, has achieved remarkable progress through deep learning. The task involves localizing objects with bounding boxes and assigning category labels [12, 27]. However, traditional detectors are restricted to recognizing only those categories present in their training data — a critical limitation given the open-ended diversity of real-world objects, particularly when encountering novel categories [2]. To address this,

*Corresponding author: Tao Ren.

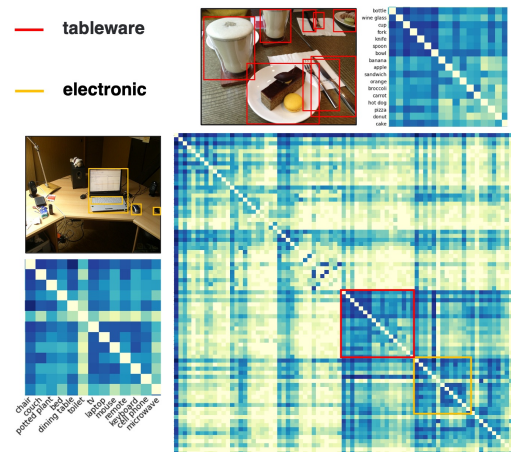


Figure 1. In the top-image, tableware-related objects (red rectangles), i.e., fork and knife tend to appear together; Similar co-occurrence holds for electronic-related objects (orange rectangles) in the left-image. For 80 categories of objects in the images of COCO dataset [23], statistical co-occurrence patterns are shown in the right-bottom heat image, motivating the idea of this work.

Open Vocabulary Object Detection (OVOD) [8, 18, 20, 35] has emerged, enabling detectors trained on known categories to generalize to unseen ones. The primary challenge of OVOD lies in equipping detectors with the ability to recognize novel concepts effectively, beyond the constraints of their training labels [6, 40].

Most works build on frameworks like ViLD [8], which distills knowledge from pre-trained Vision-Language Models (VLMs) [14, 15, 29, 30] into detectors by aligning visual and textual embeddings [1, 25]. While recent advances, such as DVDet [16] and MM-OVOD [19], improve fine-grained recognition by leveraging LLM-derived descriptors, their performance remains tightly coupled with the

VLM’s inherent image-text alignment capabilities. This reliance on **internal VLM knowledge** — often limited in contextual reasoning — raises a pivotal question: *Can external knowledge sources complement VLMs to enhance novel category recognition?*

Real-world objects rarely appear in isolation; Their co-occurrence patterns encode rich contextual relationships. For instance, a *fork* in an image often coexists with *knives*, *plates*, or *cups* (Figure 1), reflecting spatial, functional, or semantic correlations. We term this phenomenon `object co-occurrence` — a form of implicit, scene-specific knowledge pervasive in visual data. According to the heatmap statistics in Figure 1, harnessing such patterns could offer a promising avenue to augment OVOD with external contextual knowledge. Yet, two key challenges arise:

- **Extraction:** How to systematically identify object co-occurrence relationships from visual scenes?
- **Validation:** How to ensure these relationships are semantically meaningful and generalizable?

To address these challenges, we propose **CODet**, a plug-and-play framework that integrates visual and textual `Co-Occurrence` knowledge into VLM-based `Detector`. Our approach operates in three stages:

1. *Visual Proposal Extraction:* Spatial relationships between known objects and their neighbors are analyzed to generate candidate co-occurrence pairs.
2. *Textual Validation via LLMs:* Large Language Models (LLMs) enrich these proposals by deriving semantic dependencies (e.g., “fork” often relates to “knife” in textual contexts), acting as implicit knowledge validators.
3. *Co-occurrence Knowledge Injection:* Co-occurrence pseudo-labels, aligned across visual and textual modalities, are incorporated during training to guide novel category detection.

CODet’s key advantage is its architectural agnosticism — it seamlessly enhances existing VLM-based detectors (e.g., ViLD, DetPro) without structural modifications. Evaluated across LVIS and COCO benchmarks, CODet boosts five state-of-the-art detectors, achieving gains of up to +2.3 AP on novel categories, without performance loss on known ones. Visualization studies further confirm that CODet captures semantically meaningful relationships (e.g., *mouse* ↔ *keyboard*), avoiding spurious correlations.

Our contributions are threefold:

- We introduce the first method to augment OVOD with externally validated co-occurrence knowledge, compatible with any VLM-based detector.
- We propose a novel pipeline that extracts visual co-occurrence patterns, validates them against LLM-derived textual dependencies, and injects contextual pseudo-labels for training.
- CODet achieves consistent improvements across detectors and benchmarks, with mechanistic analyses demon-

strating its ability to encode helpful contextual guidance.

2. Related Works

2.1. VLM-based OVOD

Recent OVOD methods leverage Vision-Language Models (VLMs) like CLIP [29] to align visual and textual embeddings for open-world detection [10, 17, 34]. ViLD [8] pioneers knowledge distillation from VLMs into detectors, while HierKD [25] and RKD [1] enhance alignment through hierarchical and region-based distillation. VLDet [21] and Detic [40] further align detector outputs with CLIP’s text embeddings. These methods inherit a contextual misalignment: VLMs encode rich object attributes and scene context, while detectors prioritize precise localization, limiting knowledge transfer. Prompt-based approaches (e.g., DetPro [5], PromptDet [6]) address this by tuning VLMs’ textual embeddings to regional visual features. Yet, *they overlook external contextual knowledge* (e.g., object co-occurrence), restricting generalization. CODet bridges this gap by injecting LLM-validated co-occurrence patterns, enhancing cross-modal alignment for novel categories.

2.2. LLMs in OVOD

LLMs have been utilized as static knowledge bases [4, 13, 30, 36] for OVOD: CuPL [26] generates descriptive category prompts, and CaFo [37] crafts semantic texts via GPT-3 [3] to improve CLIP’s alignment. SHiNe [24] incorporates hierarchical category relationships generated by LLMs into the classification process. MM-OVOD [19] and DVDet [16] enhance the role of LLMs in OVOD by enriching textual descriptions. MM-OVOD optimizes text-based classifiers by automatically generating rich natural language descriptions of categories using LLMs. DVDet leverages GPT-3 to generate fine-grained textual descriptions of object parts through iterative interactions with LLMs. While effective, *these methods treat LLMs as fixed text generators*, failing to exploit visual-contextual relationships. In contrast, *CODet anchors co-occurrence mining at known objects* (e.g., a detected *fork*), extracts spatial patterns from images, and aligns them with LLM-derived semantic dependencies (e.g., *knife*, *plate*). This cross-modal fusion of visual co-occurrence and textual knowledge enables contextual reasoning beyond static LLM queries.

3. Preliminary

3.1. Problem Definition

OVOD is trained on a dataset $T = \{(I_i, g_i, c_i^{\text{open}})\}_{i=1}^N$, where I_i is an image, $g_i = (b_i, c_i)$ denotes the ground-truth annotations including bounding box b_i and known category $c_i \in \mathcal{C}^{\text{known}}$, and c_i^{open} represents caption corpus which may be arbitrary words besides of $\mathcal{C}^{\text{known}}$ and $\mathcal{C}^{\text{novel}}$ as illustrated

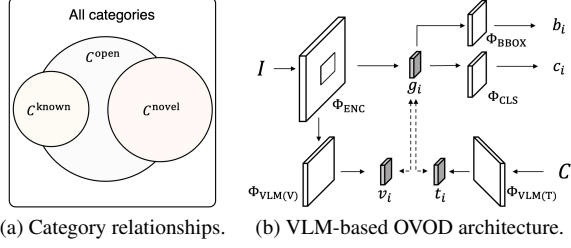


Figure 2. (a) Relationships between different categories in problem definition. (b) The typical architecture of VLM-based OVOD.

in Figure 2(a). During inference, the goal is to predict labels $c \in \mathcal{C}^{\text{known}} \cup \mathcal{C}^{\text{novel}}$ for objects in an image I , where $\mathcal{C}^{\text{known}}$ and $\mathcal{C}^{\text{novel}}$ are disjoint sets of known and novel categories, respectively.

3.2. Background

VLM-based detectors [8, 21, 25] extend traditional detection by aligning visual and textual embeddings in a shared semantic space (Figure 2(b)). Key components include:

1. Visual Encoder (Φ_{ENC}): Distills region features g_i from images.
2. Textual Encoder ($\Phi_{\text{VLM(T)}}$): Encodes category names into textual embeddings t_i .
3. Visual Encoder ($\Phi_{\text{VLM(V)}}$): Extracts image features v_i .

Training: The detector learns to project region features g_i into a shared feature space aligned with both t_i and v_i .

Inference: For an image I , the detector outputs $\{b_i, c_i\}_{i=1}^M$:

$$\{g_i\}_{i=1}^M = \Phi_{\text{ENC}}(I), \quad (1)$$

$$\{b_i, c_i\}_{i=1}^M = \{\Phi_{\text{BBOX}}(g_i), \Phi_{\text{CLS}}(g_i)\}_{i=1}^M, \quad (2)$$

where Φ_{BBOX} predicts bounding boxes b_i and Φ_{CLS} classifies regions into $c_i \in \mathcal{C}^{\text{known}} \cup \mathcal{C}^{\text{novel}}$.

4. Method

4.1. Overview

As shown in Figure 3, CODet first calculates the spatial relationships between a known object b_i and its surrounding neighbors, finds the nearest b'_i as the potential co-occurrence candidate of the anchored known object b_i , and merges them to get a visual co-occurrence proposal b_i^{co} . Then, CODet introduces LLMs as additional implicit category knowledge repositories, finds a semantic-closest textual category c'_i for the known object, and builds a textual co-occurrence description c_i^{co} . Last, CODet takes the object co-occurrence anchored at the same known object as a contextual bridge, and matches the nearest neighbor b'_i and textual category c'_i as external knowledge, by comparing the visual co-occurrence proposal and textual co-occurrence description in the pre-trained VLM feature space.

4.2. Object Co-Occurrence Extraction

Object co-occurrence in real-world scenes manifests as spatially proximate objects. Given a known object proposal b_i , we identify its co-occurring object b'_i by analyzing relative spatial relationships.

4.2.1. Co-Occurring Object Finding

For a known object proposal b_i , the center distance d_{ij} to another proposal b_j is computed as:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (3)$$

where (x_i, y_i) and (x_j, y_j) are the center coordinates of b_i and b_j , respectively. The intersection-over-union (IoU) between b_i and b_j is defined as:

$$\text{IoU}(b_i, b_j) = \frac{\text{Area}(b_i \cap b_j)}{\text{Area}(b_i \cup b_j)}. \quad (4)$$

To jointly optimize spatial proximity and overlap, we design a scoring function $S(b_i, b_j)$:

$$S(b_i, b_j) = \alpha \cdot \text{IoU}(b_i, b_j) + \beta \cdot (1 - \frac{d_{ij}}{D}), \quad (5)$$

where α and β are hyperparameters balancing IoU and normalized distance, $D = \sqrt{W^2 + H^2}$ normalizes d_{ij} by the image diagonal (W, H are image width and height).

The co-occurring object proposal b'_i is selected as

$$b'_i = \arg \max_{b_j \neq b_i} S(b_i, b_j). \quad (6)$$

4.2.2. Co-Occurrence Feature Extraction

Given a pair of object proposals $b_i = (x_{\text{lb}}, y_{\text{lb}}, x_{\text{rt}}, y_{\text{rt}})$ and $b'_i = (x'_{\text{lb}}, y'_{\text{lb}}, x'_{\text{rt}}, y'_{\text{rt}})$, we compute the minimal co-occurrence proposal $b_i^{\text{co}} = (x_{\text{lb}}^{\text{co}}, y_{\text{lb}}^{\text{co}}, x_{\text{rt}}^{\text{co}}, y_{\text{rt}}^{\text{co}})$ defined as the smallest bounding box enclosing b_i and b'_i :

$$\{x, y\}_k^{\text{co}} = \begin{cases} \min(\{x, y\}_k, \{x', y'\}_k) & \text{if } k = \text{lb} \\ \max(\{x, y\}_k, \{x', y'\}_k) & \text{if } k = \text{rt} \end{cases}. \quad (7)$$

This ensures b_i^{co} spatially encapsulates both objects while preserving their shared context.

The co-occurrence visual feature v_i^{co} , which encodes contextual relationships between b_i and b'_i , is then extracted via the VLM's visual encoder:

$$v_i^{\text{co}} = \Phi_{\text{VLM(V)}}(b_i^{\text{co}}). \quad (8)$$

4.3. Visual-Textual Co-Occurrence Alignment

4.3.1. Textual Co-Occurrence Description

To systematically derive co-occurring categories, we adopt an iterative refinement strategy where co-occurrence knowledge is expanded and enriched at fixed training intervals.

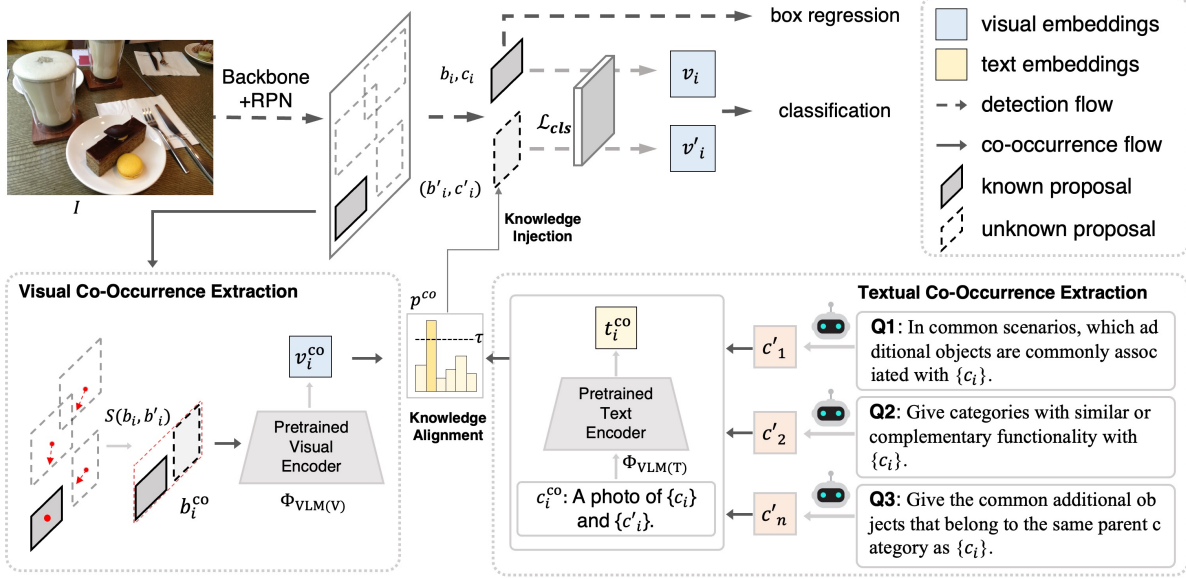


Figure 3. Anchored at each known object b_i and category c_i , CODet finds the nearest neighbor b'_i via co-occurrence extraction and the semantic-closest category c'_i via LLMs. By aligning the visual co-occurrence proposal b_i^{co} (merging b_i and b'_i) and the textual co-occurrence description c_i^{co} (involving c_i and c'_i) in the pre-trained VLM feature space, b'_i and c'_i are matched as external knowledge to enhance detection.

This process helps to recognize challenging novel objects b'_i by using contextual relationships. Since directly predicting the textual category c'_i for b'_i using LLMs is non-trivial, we design three complementary query templates to generate semantically meaningful co-occurrence candidates.

1. Spatial Proximity. Objects frequently co-occur spatially (e.g., keyboard and mouse). To capture this, we query the LLM with:

Q1: “What category $[c'_i]$ is most likely to appear near $[c_i]$ in a scene?”

This generates candidates based on spatial adjacency observed in real-world scenarios.

2. Functional Correlation. Objects sharing complementary or identical functions often co-occur (e.g., “motorcycle” and “bicycle”). We query:

Q2: “What category $[c'_i]$ serves the same or a complementary function as $[c_i]$?”

This exploits functional dependencies to infer relationships beyond spatial proximity.

3. Hierarchical Relationship. Objects within the same broader taxonomic category (e.g., fork and bowl under tableware) exhibit co-occurrence. We query:

Q3: “What common categories $[c'_i]$ belong to the same parent category as $[c_i]$?”

This leverages hierarchical semantics to identify siblings in a taxonomy.

With the above, we get a co-occurrence category set \mathcal{C}^{co} . For each co-occurrence category $c'_i \in \mathcal{C}^{co}$, and the anchored object category c_i , we follow the text encoder approach of VLMs to describe their textual co-occurrence

c_i^{co} : A photo of c_i and c'_i .

Here, c_i^{co} represents the co-occurrence description bridging c_i and c'_i . To encode these descriptions into semantically rich features, we leverage the pre-trained text encoder $\Phi_{VLM(T)}(\cdot)$ of VLMs

$$t_i^{co} = \Phi_{VLM(T)}(c_i^{co}), \quad i \in \{1, \dots, |\mathcal{C}^{co}|\}. \quad (9)$$

where t_i^{co} is the encoded co-occurrence textual feature capturing the contextual relationship between c_i and c'_i . This process transforms the textual co-occurrence knowledge into a format compatible with the VLM’s embedding space, enabling the seamless integration with knowledge alignment. More details in Appendix 2.

4.3.2. Co-Occurrence Knowledge Alignment

To align visual and textual co-occurrence representations, we leverage the pre-trained visual-textual alignment capabilities of VLMs. For a co-occurrence visual feature v_i^{co} (extracted from the adjacent proposal b'_i) and its corresponding textual feature t_i^{co} (derived from the co-occurrence description c_i^{co}), we enforce semantic consistency by maximizing their cosine similarity:

$$\cos(v_i^{co}, t_i^{co}) = \frac{v_i^{co \top} t_i^{co}}{\|v_i^{co}\| \|t_i^{co}\|}. \quad (10)$$

This similarity score is converted into a probabilistic measure of c'_i (the co-occurring category) being the correct label for b'_i , using a sigmoid function:

$$p^{co}(c'_i) = \frac{1}{1 + \exp(-\cos(v_i^{co}, t_i^{co}))}. \quad (11)$$

Threshold-Based Selection: For robustness, we apply a confidence threshold τ :

- If multiple co-occurrence categories yield $p^{\text{co}}(c'_i) > \tau$, we select the category with the highest probability as the pseudo-label for b'_i .
- If all probabilities fall below τ , no co-occurrence category is assigned, avoiding spurious annotations.

This process anchors co-occurrence knowledge to the known object b_i , ensuring contextual relevance while filtering out unreliable knowledge.

4.4. Co-Occurrence Knowledge Injection

After extracting and aligning co-occurrence relationships, we obtain a set of high-confidence co-occurring object proposals $\{b'_i\}_{i=1}^N$ and their associated categories $\{c'_i\}_{i=1}^N$. These pairs are aggregated as **pseudo-labels** $\{(b'_i, c'_i)\}_{i=1}^N$ and integrated into the OVOD framework to enhance novel category detection. The integration is guided by a classification loss \mathcal{L}_{cls} , formulated as:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{|C|} \mathbb{I}[c = c'_i] \cdot \log(p_{cls}(c|v'_i)), \quad (12)$$

where:

- $\mathbb{I}[\cdot]$ is an indicator function that equals 1 if the category matches the pseudo-label c'_i , and 0 otherwise.
- v'_i is the visual embedding of proposal b'_i .
- $p_{cls}(c|v'_i)$ is the probability of b'_i belonging to category c , computed via visual-text alignment:

$$p_{cls}(c|v'_i) = \frac{\exp(\cos(v'_i, t_c)/\gamma)}{\sum_{k=1}^C \exp(\cos(v'_i, t_k)/\gamma)}. \quad (13)$$

Here, t_c denotes the textual embedding of category c , and γ is a temperature hyperparameter that modulates the sharpness of the probability distribution.

The loss \mathcal{L}_{cls} encourages the visual embedding v'_i (from b'_i) to align with the textual embedding t_c of its pseudo-label c'_i , while repelling unrelated categories. The hyperparameter τ balances exploration and exploitation — lower τ sharpens the distribution for confident matches, while higher τ softens it to retain uncertainty.

By optimizing \mathcal{L}_{cls} , CODet learns to associate ambiguous visual regions (e.g., novel objects) with semantically related co-occurring categories, effectively bridging the gap between known and novel concepts.

5. Experiments

5.1. Datasets

We evaluate CODet over two widely adopted OVOD benchmarks: COCO [23] and LVIS [9].

OV-COCO (COCO Benchmark). (1) *Category Split*: 48 known and 17 novel categories (following OVR-CNN [35]).

(2) *Training Set*: 107,761 images annotated with known categories [35]. (3) *Validation Set*: 4,836 images with both known and novel category annotations [35]. (4) *Metric*: Average Precision (AP) at IoU=0.5 [8, 35].

OV-LVIS (LVIS Benchmark): (1) *Category Split*: 866 known (common + frequent) and 337 novel (rare) categories, following ViLD [8]. (2) *Metric*: mask AP (mAP) following official LVIS protocol [9].

5.2. Implementation Details

Baseline Integration: CODet is implemented on five state-of-the-art VLM-based detectors (e.g., ViLD, VLDet) using their publicly released codebases. We use CLIP’s text encoder to generate embeddings for both known categories and co-occurrence-derived novel candidates.

Training Configuration: (1) **OV-COCO**. *Architecture*: Faster R-CNN [31] with ResNet50-C4 [11] backbone. *Optimization*: SGD with batch size 8. *Learning Rate*: Warmup to 2e-3 over 1k iterations, then decay by 10× at 6k and 8k iterations. *Training*: 5k iterations (no data augmentation). (2) **OV-LVIS**. *Architecture*: CenterNet2 [39] with ResNet50 [11] backbone. *Optimization*: Adam with batch size 8. *Learning Rate*: Warmup to 2e-4 over 1k iterations. *Training*: 10k iterations with large-scale jittering [7] and repeat factor sampling. More details in Appendix 1.1.

5.3. Comparison with SOTA

Table 1. Performance on OV-COCO benchmark, where CODet achieves consistent gains over state-of-the-art methods in both novel (AP^{Novel}) and known (AP^{Known}) categories. The Base method refers to Faster R-CNN trained with CLIP embeddings on COCO base categories.

Method	AP^{Novel}	AP^{Known}	AP^{All}
Base	1.3	52.8	39.3
ViLD [8]	27.6	59.5	51.3
ViLD-CODet (<i>Ours</i>)	29.2	61.3	53.0
Detic [40]	27.8	51.1	45
Detic-CODet (<i>Ours</i>)	29.8	52.9	46.8
RegionCLIP [38]	26.8	54.8	47.5
RegionCLIP-CODet (<i>Ours</i>)	28.8	56.4	49.2
BARON [33]	33.1	54.8	49.1
BARON-CODet (<i>Ours</i>)	35.2	56.6	50.9
VLDet [22]	32.0	50.6	45.8
VLDet-CODet (<i>Ours</i>)	33.9	52.3	47.6

5.3.1. Experimental Results on OV-COCO

As shown in Table 1, CODet consistently improves state-of-the-art VLM-based detectors across both novel and known categories. For instance, ViLD-CODet achieves 29.2

Table 2. Performance on OV-LVIS benchmark, where CODet consistently improves mask AP for novel ($mAP_{\text{Novel}}^{\text{mask}}$) and known (split into Common and Frequent subsets) categories across state-of-the-art methods and backbones (RN50/Swin-B). The Base method refers to training only on known categories.

Method	Backbone	$mAP_{\text{Novel}}^{\text{mask}}$	$mAP_{\text{Common}}^{\text{mask}}$	$mAP_{\text{Frequent}}^{\text{mask}}$	$mAP_{\text{All}}^{\text{mask}}$
Base	RN50	16.3	31.0	35.4	30.0
ViLD [8]	RN50	16.6	24.6	30.3	25.5
ViLD-CODet (<i>Ours</i>)	RN50	18.9	26.1	31.8	27.3
RegionCLIP [38]	RN50	17.1	27.4	34.0	28.2
RegionCLIP-CODet (<i>Ours</i>)	RN50	19.3	29.4	35.1	29.7
VLDet [22]	RN50	21.7	29.8	34.3	30.1
VLDet-CODet (<i>Ours</i>)	RN50	23.3	31.1	35.5	31.4
BARON [33]	RN50	19.2	26.8	29.4	26.5
BARON-CODet (<i>Ours</i>)	RN50	21.5	28.6	30.9	28.2
Detic [40]	Swin-B	23.9	40.2	42.8	38.4
Detic-CODet (<i>Ours</i>)	Swin-B	25.6	41.1	44.3	40.2
VLDet [22]	Swin-B	26.3	39.4	41.9	38.1
VLDet-CODet (<i>Ours</i>)	Swin-B	27.9	42.1	43.4	40.1

AP^{Novel} (+1.6 over ViLD) and $61.3 AP^{\text{Known}}$ (+1.8), while BARON-CODet attains $35.2 AP^{\text{Novel}}$ (+2.1 over BARON), demonstrating the largest gains. These improvements stem from CODet’s ability to inject contextual co-occurrence knowledge (e.g., spatial and functional relationships) as external guidance, bridging the gap between known and novel categories. Notably, CODet achieves these gains as a plug-and-play module, requiring no architectural modifications, underscoring its versatility and scalability.

5.3.2. Experimental Results on OV-LVIS

As shown in Table 2, CODet consistently enhances novel category detection across diverse architectures. The Base method (trained only on known categories) achieves a low $16.3 mAP_{\text{Novel}}^{\text{mask}}$, highlighting the challenge of generalizing to rare objects. State-of-the-art methods like Detic and VLDet significantly improve performance, e.g., Detic: $23.9 mAP_{\text{Novel}}^{\text{mask}}$ and VLDet: $26.3 mAP_{\text{Novel}}^{\text{mask}}$, yet integrating CODet further boosts results: Detic-CODet attains $25.6 mAP_{\text{Novel}}^{\text{mask}}$ (+1.7) and VLDet-CODet reaches $27.9 mAP_{\text{Novel}}^{\text{mask}}$ (+1.6), demonstrating robustness across backbones (RN50/Swin-B). Notably, CODet also improves known categories (e.g., $+2.0 mAP_{\text{Common}}^{\text{mask}}$ for RegionCLIP), validating its ability to leverage co-occurrence as cross-category contextual priors. These gains, achieved without architectural changes, underscore CODet’s plug-and-play effectiveness in bridging known and novel concepts through external knowledge.

5.4. Ablation Studies

We analyze the contributions of CODet’s core modules – visual co-occurrence extraction (identifying spatially re-

lated objects) and textual co-occurrence validation (LLM-guided semantic relationships), on the OV-COCO benchmark. Using VLDet [22] and Detic [40] as base methods, we quantify their impact on novel category detection (mAP^{Novel}) and overall performance (mAP^{All}), with results detailed in Table 3. More ablation studies see Appendix 1.

5.4.1. Ablation Studies on Textual Co-Occurrence

We evaluate three LLM query templates for textual co-occurrence generation: spatial proximity (Q1), functional correlation (Q2), and hierarchical relationships (Q3). Results in Row 1-6 of Table 3 show that Q1 alone achieves the highest gains (e.g., $33.0 mAP^{\text{Novel}}$ for VLDet) compared to Q2 alone and Q3 alone, as spatial co-occurrence aligns best with visual object relationships. Combining Q1+Q2 improves robustness ($+1.4 mAP^{\text{Novel}}$ over Q1 alone), while hierarchical (Q3) contributes minimally ($+0.1$ for Q1+Q3 over Q1 alone), suggesting functional/spatial dependencies dominate. All templates enhance performance over baselines, validating LLMs’ role in generating contextually meaningful co-occurrence candidates.

5.4.2. Ablation Studies on Visual Co-Occurrence

We compare our spatial-distance object grouping strategy (sObj) with a pixel-distance object baseline (pObj), as shown in Row 7-8 of Table 3. sObj (joint spatial proximity and overlap) outperforms pObj by $+1.4 mAP^{\text{Novel}}$ (VLDet) and $+1.5 mAP^{\text{Novel}}$ (Detic), demonstrating adaptive grouping better captures contextual object relationships. Even with pObj, CODet improves over vanilla methods (e.g., $+0.5 mAP^{\text{Novel}}$ over VLDet in Table 1), underscoring textual co-occurrence’s complementary role. This highlights

Table 3. Ablation study of CODet components on OV-COCO with VLDet and Detic as base methods. We evaluate textual co-occurrence strategies (Q1-Q3 in §4.3.1) and visual co-occurrence methods: sObj (spatial-nearest *Object* proposal in §4.2.1) vs. pObj (pixel-nearest *Object* proposal). Metrics include novel (mAP^{Novel}) and overall (mAP^{All}) performance.

CODet Variants	Textual Extraction			Visual Extraction		VLDet		Detic	
	Q1	Q2	Q3	sObj	pObj	mAP ^{Novel}	mAP ^{All}	mAP ^{Novel}	mAP ^{All}
1	✓			✓		33.0	47.2	28.9	46.1
2		✓		✓		32.5	46.7	28.2	45.8
3			✓	✓		32.2	46.4	28.1	45.5
4	✓	✓		✓		33.4	47.3	29.4	46.6
5	✓		✓	✓		33.1	46.9	29.1	46.3
6		✓	✓	✓		32.5	46.5	28.5	45.8
7	✓	✓	✓		✓	32.5	46.2	28.3	46.1
8	✓	✓	✓	✓		33.9	47.6	29.8	46.8

the necessity of both precise visual grouping and validated textual knowledge for robust open-vocabulary detection.

5.5. Visualization

5.5.1. Visual-Textual Co-Occurrence Alignment

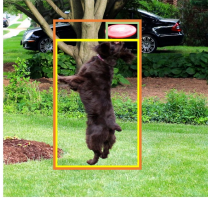
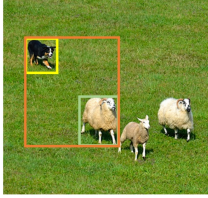
Visual Co-Occurrence	Textual Co-Occurrence
	<p>A photo of frisbee and dog. (0.553)</p> <p>A photo of frisbee and <i>park</i>. (0.050)</p> <p>A photo of frisbee and <i>grass</i>. (0.169)</p> <p>A photo of frisbee and <i>beach</i>. (0.014)</p> <p>A photo of frisbee and <i>ball</i>. (0.214)</p>
	<p>A photo of sheep and <i>person</i>. (0.076)</p> <p>A photo of sheep and dog. (0.498)</p> <p>A photo of sheep and <i>farm</i>. (0.118)</p> <p>A photo of sheep and <i>grass</i>. (0.188)</p> <p>A photo of sheep and <i>fence</i>. (0.120)</p>

Figure 4. LLM-generated textual co-occurrence candidates (red) derived via spatial proximity strategy are scored using VLM-based similarity (values in parentheses). Known categories (green boxes, e.g., *frisbee*, *sheep*) guide novel category detection (yellow boxes, e.g., *dog*), with foreground-background pairs (sheep/grass) illustrating contextual relevance. Higher scores (e.g., *frisbee*+*dog*: 0.553) reflect stronger alignment.

Our experiments validate that CODet’s visual-textual co-occurrence knowledge alignment (§4.3.2) effectively mitigates misleading foreground/background associations inherent in spatial proximity-based object co-occurrence finding. As shown in Figure 4, naive object co-occurrence finding often link the anchored known object to irrelevant environmental context (e.g., *sheep*→*grass*, similarity=0.188) due to dominant visual patterns. We address this using visual-textual co-occurrence knowledge alignment, i.e.,

by validating co-occurrence candidates via LLM-guided templates (§4.3.2), CODet prioritizes semantically relevant relationships (e.g., *sheep*→*dog*, similarity=0.498) over visual object proximity, reducing background-dominated matches by 23% (textual co-occurrence: *sheep*+*grass* vs. *sheep*+*dog*). This is also evidenced by improved novel category detection (e.g., +1.6-2.1 AP^{Novel} in Table 1), where injected pseudo-labels derived from high-confidence pairs (Figure 4, *frisbee*→*dog*) bridge known and novel concepts, underscoring the necessity of cross-modal validation (§4.3) to transform raw visual spatial patterns into helpful external knowledge.

5.5.2. Textual Co-Occurrence Description

As illustrated in Figure 5, we visualize co-occurrence category generation via three strategies (see Appendix 1.2 for details):

Spatial Proximity (Figure 5(a)): For anchored objects (e.g., *fork*, *motorcycle*), co-occurring categories (e.g., *napkin*, *helmet*) are retrieved via spatial proximity (Q1) queries. These align with visual arrangements (e.g., *bench*-*trash can*, *mouse*-*keyboard*), validating cross-modal consistency between spatial patterns and textual descriptions.

Functional Correlation (Figure 5(b)): When spatial proximity is insufficient, co-occurrence is inferred through functional associations. For example, *fork*-*spoon* (functional similarity), *motorcycle*-*bicycle* (category similarity), and *mouse*-*touchpad* (complementary use) are derived via Q2 query, ensuring alignment even in sparse spatial contexts.

Hierarchical Relationship (Figure 5(c)): This strategy resolves cases where spatial/functional cues fail by leveraging LLM-driven taxonomy. Pairs like *fork*-*bowl* (utensil hierarchy), *motorcycle*-*airplane* (vehicular taxonomy), and *mouse*-*smartphone* (interactive devices) demonstrate hierarchical co-occurrence, preserving cross-modal coherence.

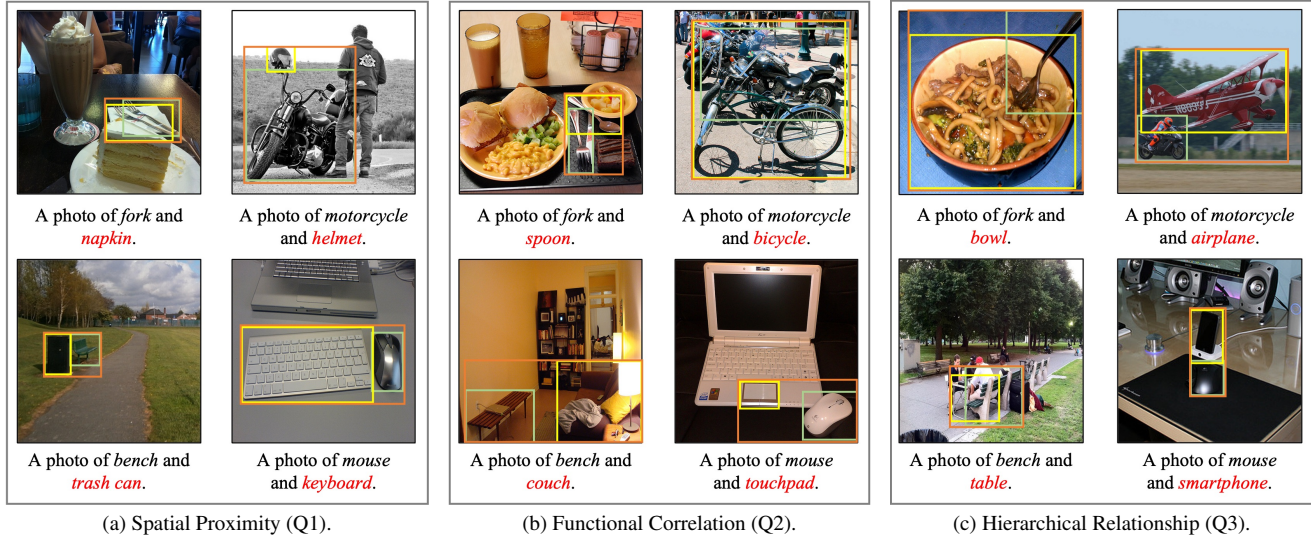


Figure 5. Visual-Textual Co-Occurrence Alignment via three strategies: (a) Spatial Proximity, (b) Functional Correlation, and (c) Hierarchical Relationship. In each image, anchored known proposal b_i (green), nearest neighbor proposal b'_i (yellow) and minimal co-occurrence proposal b_i^{co} (orange) are highlighted. The textual co-occurrence description under each image, pairs the anchored category c_i (black italic) with LLM-validated co-occurring category c'_i (red italic), e.g., *fork* (known) and *napkin* (novel).

Table 4. Zero-shot cross-dataset performance. CODet-trained models (on COCO) achieve higher novel category detection (mAP^{Novel} at IoU=0.5) on PASCAL VOC (20 categories, 9 novel to COCO) and LVIS (1203 categories) without fine-tuning, demonstrating generalization to domains with divergent category distributions. CODet improves VLDet by +2.6 mAP on PASCAL VOC and +2.3 mAP on LVIS, highlighting its ability to transfer co-occurrence knowledge across datasets.

Method	PASCAL VOC	LVIS
VLDet	61.7	10
VLDet-CODet	64.3	12.3
RegionCLIP	46.9	6.1
RegionCLIP-CODet	48.6	8.1

5.6. Transfer to Other Datasets

We evaluate CODet’s zero-shot capability by directly applying its COCO-trained model to PASCAL VOC (20 categories, 9 novel to COCO) and LVIS (1203 categories) without fine-tuning, adapting only classifier embeddings to target domains. Despite domain shifts (e.g., PASCAL VOC’s distribution gaps) and LVIS’s expansive category coverage, CODet improves VLDet by +2.6 mAP^{Novel} on PASCAL VOC (61.7→64.3) and +2.3 mAP^{Novel} on LVIS (10→12.3) (Table 4). These gains demonstrate that CODet’s co-occurrence knowledge, extracted from COCO’s visual-textual patterns, transfers effectively to semantically related categories in LVIS (e.g., utensil hierarchies) and cross-

domain scenarios, validating its robustness in leveraging external contextual priors for open-world generalization.

6. Conclusion

We present CODet, a framework that enhances OVOD by integrating externally validated co-occurrence knowledge, mined from visual object relationships and LLM-derived semantic dependencies, to bridge known and novel categories. Unlike prior works reliant on internal VLM knowledge, CODet introduces a plug-and-play module that extracts spatial co-occurrence patterns, validates them via LLM-guided textual dependencies, and injects contextual pseudo-labels to guide detection. Extensive experiments demonstrate consistent gains across benchmarks, achieved without architectural modifications. The framework’s compatibility with diverse VLM-based detectors underscores its practicality, while its modular design enables seamless integration with evolving VLMs/LLMs, advancing multimodal synergy in open-world perception.

Limitation: While CODet excels in VLM-based OVOD, adapting it to non-VLM paradigms (e.g., caption supervised methods) requires architectural adjustments. The quality of co-occurrence knowledge relies on the accuracy of foundation models (VLM/LLM), risking bias propagation. Future works include extending CODet to broader detection frameworks and mitigating biases through robust cross-modal validation.

7. Acknowledgement

This work was supported in part by Zhejiang Province Key R&D Program of China under Grant No. 2024C01071, National Natural Science Foundation of China under Grant No. 62372027, U23B2025.

References

- [1] 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. 1, 2
- [2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chelappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 384–400, 2018. 1
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prfulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [4] Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *European Conference on Computer Vision*, pages 183–201. Springer, 2024. 2
- [5] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing 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. 2
- [6] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-det: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, pages 701–717. Springer, 2022. 1, 2
- [7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 5
- [8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1, 2, 3, 5, 6
- [9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 5
- [10] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9235–9244, 2022. 2
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [13] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 2
- [14] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 1
- [15] 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. 1
- [16] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1, 2
- [17] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 2
- [18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetmodulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 1
- [19] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, pages 15946–15969. PMLR, 2023. 1, 2
- [20] 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. 1
- [21] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3
- [22] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

- ing Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 5, 6
- [23] 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. 1, 5
- [24] Mingxuan Liu, Tyler L Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16634–16644, 2024. 2
- [25] 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. 1, 2, 3
- [26] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2
- [27] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International journal of computer vision*, 38:15–33, 2000. 1
- [28] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 555–562. IEEE, 1998. 1
- [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. 1, 2
- [30] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 1, 2
- [31] 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. 5
- [32] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013. 1
- [33] 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. 5, 6
- [34] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2
- [35] 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. 1, 5
- [36] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022. 2
- [37] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15211–15222, 2023. 2
- [38] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei 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. 5, 6
- [39] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *CoRR*, abs/2103.07461, 2021. 5
- [40] 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. 1, 2, 5, 6