NoOVD: Novel Category Discovery and Embedding for Open-Vocabulary Object Detection

Yupeng Zhang[1,2] Ruize Han[3] Zhiwei Chen[4] Wei Feng[1,2] Liang Wan[1,2*]
[1]College of Intelligence and Computing, Tianjin University.
[2]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage.
[3]Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology.
[4]School of Artificial Intelligence, Nanchang University.
{zhangyupeng, wfeng, lwan}@tju.edu.cn, hanruize@suat-sz.edu.cn, zhiweichen@ncu.edu.cn

Abstract

Despite the remarkable progress in open-vocabulary object detection (OVD), a significant gap remains between the training and testing phases. During training, the RPN and RoI heads often misclassify unlabeled novel-category objects as background, causing some proposals to be prematurely filtered out by the RPN while others are further misclassified by the RoI head. During testing, these proposals again receive low scores and are removed in post-processing, leading to a significant drop in recall and ultimately weakening novel-category detection performance. To address these issues, we propose a novel training framework—NoOVD—which innovatively integrates a self-distillation mechanism grounded in the knowledge of frozen vision-language models (VLMs). Specifically, we design K-FPN, which leverages the pretrained knowledge of VLMs to guide the model in discovering novel-category objects and facilitates knowledge distillation—without requiring additional data—thus preventing forced alignment of novel objects with background. Additionally, we introduce R-RPN, which adjusts the confidence scores of proposals during inference to improve the recall of novel-category objects. Cross-dataset evaluations on OV-LVIS, OV-COCO, and Objects365 demonstrate that our approach consistently achieves superior performance across multiple metrics.

1. Introduction

General object detection techniques have achieved remarkable progress driven by deep neural networks, with methods such as Faster R-CNN [26] delivering outstanding performance. However, their training still heavily depends on extensive manual annotation, requiring tens of thousands of bounding boxes per category—an expensive and inefficient process. Although high-quality datasets like Pascal VOC [5] and MS COCO [19] are available, the limited number of categories they cover falls far short of human cognitive capabilities. Achieving universal object detection demands exponentially increasing resources, presenting a formidable challenge.

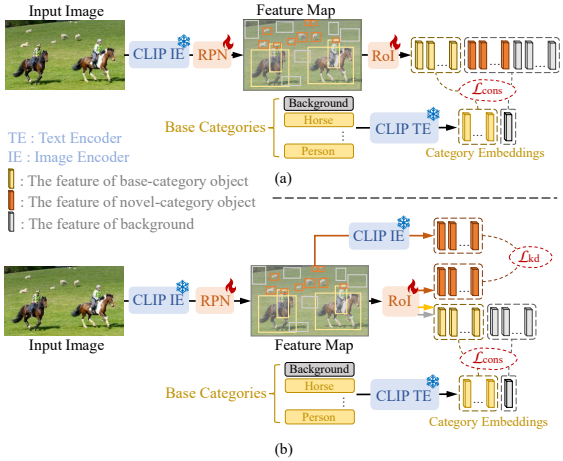

Figure 1. Training process for OVD using frozen CLIP. (a) Commonly used training process, (b) Our training process.

In recent years, large-scale vision-language models (VLMs) such as CLIP [24] and ALIGN [10] demonstrate exceptional zero-shot classification capabilities, driving significant advancements in computer vision. Open-vocabulary object detection (OVD) emerges to overcome the closed-set limitations of traditional detectors. Leading approaches [1, 4, 7, 18, 21] rely on VLMs, transferring their zero-shot abilities to detectors through knowledge distillation or image-text embedding alignment. Other methods [15, 15, 20, 31, 32, 35, 38] construct detectors based on frozen VLMs, avoiding knowledge degradation during distillation or fine-tuning and preserving generalization to the greatest extent. Although these methods achieve impressive

*Corresponding author.

performance in novel category detection, a considerable gap still exists between the training and testing categories.

In the two-stage detection framework built upon a frozen VLM (*e.g.*, CLIP), as shown in Fig. 1 (a), the model is trained using only labeled data from base categories. During the Region Proposal Network (RPN) training phase, all latent novel-category objects are forcibly regarded as background. In the classification phase, the model similarly forces the alignment of novel-category object features with the background text embeddings, which severely hinders the knowledge transfer capabilities of VLMs. This way, during testing, the proposals for novel-category objects generated by the RPN receive low scores due to being misclassified as background, resulting in their filtration during the RPN's post-processing stage. This significantly reduces the detection recall of novel-category objects. These limitations ultimately impair detection performance, particularly affecting the recognition of novel-category objects. Previous methods typically rely on large-scale data training [2, 13, 30, 36] or pseudo-labeling strategies [1, 9, 17, 31, 42, 45] to discover novel-category objects. However, large-scale data collection and training incur substantial costs, while pseudo-labeling methods depend on text matching and inevitably introduce noise, resulting in limited generalization to novel categories.

As shown in Fig. 1(b), we revisit the OVD training pipeline built on frozen VLMs and introduce an entirely new detection framework. During training, under supervision from base categories, we first propose the learnable-parameter-free knowledge-retentive FPN (K-FPN), which builds a knowledge-retentive feature pyramid directly from the frozen multi-layer CLIP features, thereby maximizing the retention of CLIP's representation capacity for novel categories. We then project RPN proposals onto K-FPN and, using diverse foreground/background text descriptions generated by an LLM together with CLIP's zero-shot recognition ability, identify latent novel-category proposals and inject their knowledge into the detector through self-distillation—effectively preventing novel-category features from being forced to align with background. During testing, to prevent novel-category proposals from being prematurely filtered due to low RPN scores, we reuse the training-stage discovery strategy to identify latent novel objects and fuse their foreground scores with the original RPN confidence. The updated proposals are then re-ranked, and the top-K are fed into the RoI head, substantially improving the recall of novel categories. Notably, unlike previous methods, the entire process requires no additional training data and does not rely on constructing pseudo image–text pairs, thereby eliminating pseudo-label noise.

In summary, the main contributions of this work are:

- We propose a novel OVD framework that focuses on the discovery of novel categories. The model simultaneously learns base-category knowledge, identifies latent novel objects, and performs knowledge self-distillation, effectively avoiding the misclassification of novel categories as background and preserving the novel-category knowledge embedded in VLMs.
- We propose a knowledge-retentive FPN (K-FPN) built on frozen CLIP to preserve world knowledge for novel-category discovery and self-distillation. During testing, we add a Re-weighted RPN (R-RPN) to improve the recall of latent novel-category objects.
- Comprehensive evaluations on both OV-LVIS [8] and OV-COCO [19] benchmarks, coupled with cross-dataset validation on Objects365 [28] validation set, consistently demonstrate the SOTA performance, which conclusively establishes the superior robustness and effectiveness of the proposed OVD framework.

2. Related Work

Open-Vocabulary Object Detection (OVD). With the rapid advancement of vision–language models (VLMs) such as CLIP [24] and ALIGN [10], open-vocabulary object detection (OVD) becomes an important research direction, enabling models to recognize both base and novel categories within a unified cross-modal semantic space. Existing approaches largely rely on region–text pairs, knowledge distillation, transfer learning, or pseudo-labeling paradigms: ViLD [7] and DetPro [4] distill CLIP knowledge into detectors; OADP [33] and DK-DETR [18] strengthen semantic modeling during distillation; RO-ViT [13], CORA [36], YOLO-World [2], and YOLOE [30] train on large-scale region–text datasets to achieve efficient open-world perception, **albeit at high cost.** Pseudo-labeling methods such as Detic [45], OCO [1], Proxy-Det [9], LBP [17], SAS-Det [42], and OV-DQUO [31] mine potential objects using image-level tags, class-agnostic detectors, or proxy categories; **however, their reliance on text matching inevitably introduces noise, resulting in limited generalization to novel categories.** Meanwhile, methods like F-VLM [15], CLIPSelf [35], DST-Det [38], and DeCLIP [32] build two-stage detectors on frozen CLIP models, training only the detection heads to achieve open-vocabulary recognition. **Yet, most OVD methods mistakenly treat novel-category objects as background during training and as foreground during inference, leading to a significant training–inference mismatch.**

To address these issues, we leverage the zero-shot capability of frozen VLMs during training to proactively identify latent novel-category objects and introduce a self-distillation mechanism that prevents them from being aligned with background, thereby eliminating noise introduced by pseudo image–text pairs at the source. Notably, our method requires no additional data or changes to the training pipeline while substantially improving novel-

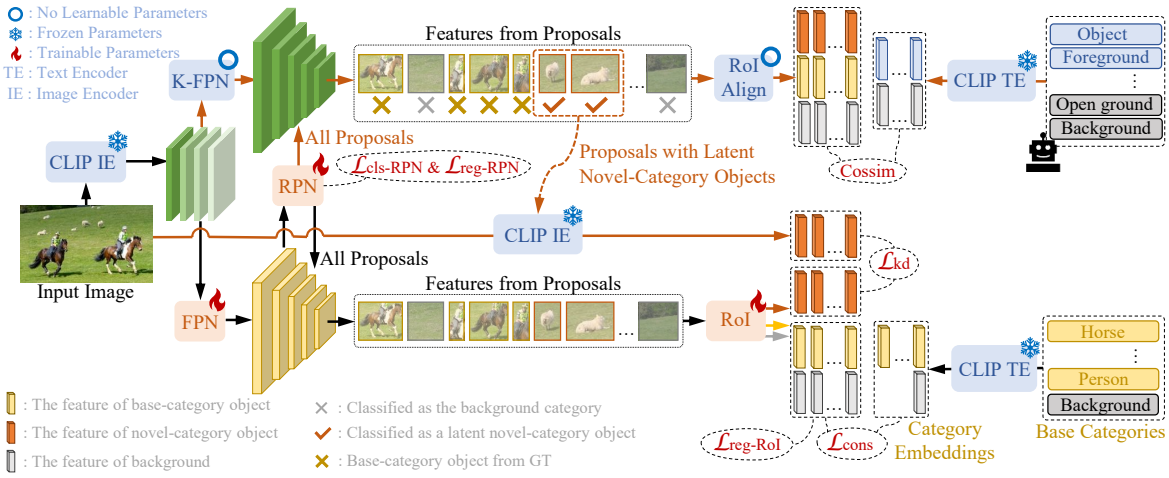

Figure 2. Illustration of the training process of NoOVD. We use K-FPN to extract the pyramid embeddings from frozen CLIP to identify latent novel-category objects. Besides the image-text alignment with $\mathcal{L}_{\text{cons}}$, we also align the features of the RoI head with the features from K-FPN via knowledge self-distillation with $\mathcal{L}_{\text{kd}}$.

category detection performance.

Vision-Language Models (VLMs). Recent years have witnessed groundbreaking advances in VLMs through contrastive learning. Pioneering works such as CLIP [24] and ALIGN [10] establish unified image-text feature spaces by learning cross-modal alignment on web-scale datasets. These models demonstrate remarkable zero-shot transfer capabilities—for instance, CLIP achieves open-vocabulary classification on ImageNet [3] without fine-tuning. This paradigm inspires numerous innovations [4, 6, 7, 12, 15, 16, 18, 22, 23, 25, 32–35, 37, 40, 41, 44] for downstream tasks including segmentation and object detection. We present an OVD framework that efficiently transfers CLIP's pretrained knowledge. During training, the CLIP encoders are frozen, and their cross-modal alignment is leveraged to uncover latent novel-category objects. A self-distillation mechanism then guides the detector to learn such knowledge, enabling effective transfer while maintaining computational efficiency.

3. Methodology

3.1. Preliminaries and Framework

Given a target image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ as input to the detector, two types of outputs are typically desired: (1) Location, where the bounding box coordinates $\mathbf{b}_i \in \mathbb{R}^4$ represent the location of the i-th predicted object. (2) Category, where $P_i^j \in \mathcal{C}_{\text{test}}$ representing the j-th category is assigned to $\mathbf{b}_i$, with $\mathcal{C}_{\text{test}}$ as the set of categories during testing. Following the open-set setting, in the training phase, the categories consist solely of the set of base category $\mathcal{C}_{\text{base}}$, In the testing phase, the vocabulary is extended to include the set of novel category names $\mathcal{C}_{\text{novel}}$, $\mathcal{C}_{\text{test}} = \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}}$ with $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$.

Overview. Our core idea is to fully exploit the strong generalization capabilities of VLMs, *i.e.* CLIP, which arise from extensive pretraining on large-scale image–text pairs. As shown in Fig. 2, we use frozen CLIP image and text encoders as the backbone and train only the detection modules (FPN, RPN, and RoI) with base-category labels. This structure not only preserves CLIP's pre-trained knowledge but also significantly reduces training costs.

We propose a novel detection framework. During training, we build the detector under supervision from base categories and introduce a learnable-parameter-free architecture, K-FPN, which directly constructs a feature pyramid from frozen multi-layer CLIP features, thereby preserving CLIP's pretrained knowledge to the greatest extent. We then combine diverse foreground–background textual prompts to leverage CLIP's zero-shot capability for discovering latent novel-category objects and integrate this knowledge into the detector through self-distillation. During testing, we introduce a novel-category retention strategy that identifies latent novel-category objects before RPN post-processing and boosts the confidence scores of their proposals, thereby significantly improving novel-category recall.

3.2. K-FPN for Knowledge Retention

To address the limitation of existing methods that implicitly treat latent novel-category objects as background during training, we propose a new method that simultaneously learns from base-category labels and leverages CLIP's pretrained knowledge. However, as the model is trained with only limited base-category data, the features extracted by frozen CLIP tend to drift when passing through subsequent modules with learnable parameters (*e.g.*, FPN), making it difficult to fully preserve CLIP's knowledge. This makes it difficult to detect latent novel-category objects through the generated feature pyramid. Therefore, a core issue is how to retain the knowledge for novel-category objects discovery.

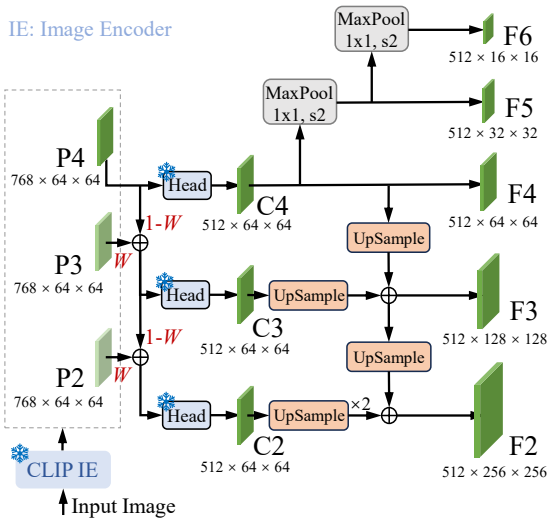

Figure 3. Overall of K-FPN (the CLIP Image Encoder is taken as an example with ViT-B/16).

To address this, we propose a novel *learnable-parameter-free* network, i.e., K-FPN (Knowledge-retentive FPN), that directly utilizes the frozen multi-layer feature maps of CLIP to construct a hierarchical feature pyramid, aiming to preserve CLIP's original knowledge as much as possible. This enables us to effectively discover latent novel-category objects through textual information. The overall architecture is shown in Fig. 3.

Taking CLIPSelf ViT-B/16 as an example, we first extract image features using CLIP's Image encoder, as

$$F_I = \mathit{CLIP\ IE}(\mathit{Image}), \tag{1}$$

where *CLIP IE* represents CLIP's Image Encoder, *Image* refers to the input image, and F_I represents the feature of the input image.

Following the detection framework F-ViT (proposed in CLIPSelf [35]), we select the feature maps from layers [5, 7, 11] of F_i as the base feature maps, denoted as {P2, P3, P4}, each with a resolution of 768 × 64 × 64. To ensure that each layer retains rich semantic information, we perform top-down feature fusion in an FPN manner. We then employ the **dimensionality-reduction head** used in the self-distilled CLIP from CLIPSelf, freeze its parameters, and convert the channel dimension of the fused features from 768 to 512 (aligned with the text embedding dimension), yielding the more detail-preserving feature maps {C2}, {C3}, and {C4}.

Then we upsample {C4} — containing higher-level semantic information — using bilinear interpolation, a standard feature-processing operation, and subsequently concatenate it laterally with the upsampled {C3} and {C2} (which, though lower level, provide more precise localization), producing higher-resolution feature maps {F2, F3, F4}. Finally, by applying max pooling twice to {C4}, consistent with the FPN design, we obtain {F5, F6}. This yields a five-level feature pyramid {F2, F3, F4, F5, F6}, each layer having 512 channels and sharing the same spatial resolutions as the original FPN. We represent this process as

$$F_{\textit{K-FPN}} = \textit{K-FPN}(F_I), \tag{2}$$

where *K-FPN* represents the process of K-FPN, and $F_{\textit{K-FPN}}$ denotes the feature pyramid generated by K-FPN. Notably, the entire process involves no learnable parameters, thus maximizing the preservation of CLIP's knowledge, providing a solid foundation for foreground object discovery.

3.3. Novel Category Discovery and Embedding

Based on the features from K-FPN retaining the knowledge of CLIP, we then consider how to leverage the zero-shot recognition capability of pre-trained large models to uncover latent novel-category objects. It prevents the misalignment of novel-category features with background semantics, thereby mitigating semantic bias. After that, we perform knowledge self-distillation by aligning RoI features with CLIP features, enabling effective knowledge transfer. This process has two main components:

General description guided novel category discovery. Given a text prompt, CLIP exhibits impressive zero-shot recognition abilities. However, to effectively identify all foreground objects (both base and novel categories) during the RPN phase, it is necessary to accurately input all category names, which is impractical in real-world applications. Therefore, we design a class-agnostic latent novel-category object discovery strategy, which consists of two steps:

① *Foreground-background text descriptions.* To maximize the detection of all foreground objects during the RPN stage, we explore the design of text prompts for foreground objects that are agnostic to specific categories, with the goal of enhancing foreground object detection recall. We use ChatGPT-o1 to generate diverse foreground object descriptions as prompts, aiming to detect all foreground objects rather than focusing on any specific category. We use 'object' as the base noun and combine it with higher-level semantic terms (such as 'plant' and 'animal') to cover a broader range of object categories. The final format is exemplified as: '`This is an object, specifically a plant`' or '`This is an object, specifically an animal`'. To maintain balance with the number of foreground category descriptions, we also provide various background descriptions, such as '`This is a background area`' and '`This is part of the ground`'. Then, we extract the embedding using the frozen CLIP text encoder,

$$E_T = \mathit{CLIP\ TE}(\mathit{Text}), \tag{3}$$

where *CLIP TE* represents CLIP's Text Encoder, *Text* refers to the foreground-background text descriptions, and E_T represents the embeddings of the text descriptions.

② *Novel-category object discovery.* We directly map the proposals generated by the RPN to K-FPN, crop their features, and extract the proposal features using RoI Align, as

$$F^{K\text{-}FPN}_{proposals} = RoI\ Align\left(Crop_{proposals}\left(F_{K\text{-}FPN}\right)\right), \quad (4)$$

where $Crop_{proposals}$ denotes the cropping operation based on proposals, and $F^{K\text{-}FPN}_{proposals}$ represents the features obtained by cropping $F_{K\text{-}FPN}$ based on proposals. We then compute the cosine similarity s between these features $F^{K\text{-}FPN}_{proposals}$ and category-agnostic foreground-background text enbeddings E_T, as

$$s = \frac{F^{K\text{-}FPN}_{proposals} \cdot E_T}{\|F^{K\text{-}FPN}_{proposals}\| \cdot \|E_T\|}. \quad (5)$$

We retain the proposals belonging to the foreground and discard those classified as base categories based on the GT bbox. The remaining proposals are considered to contain latent novel-category objects.

Novel-category object embedding via knowledge self-distillation. After selecting these proposals, we crop the corresponding regions from the original image and pass them through the frozen CLIP model to extract features, as

$$F^{Image}_{proposals+} = CLIP\ IE\left(Crop_{proposals+}\left(Image\right)\right), \quad (6)$$

where *proposals+* denotes the proposals containing latent novel-category objects, and $F^{Image}_{proposals+}$ represents the features of the latent novel-category objects extracted by the frozen CLIP Image Encoder. Subsequently, the features extracted by the frozen CLIP are aligned with the proposal features obtained through RoI, as

$$F^{RoI}_{proposals+} = RoI\ Align\left(Crop_{proposals+}\left(F_{FPN}\right)\right), \quad (7)$$

$$\mathcal{L}_{\text{kd}} = \|F^{RoI}_{proposals+} - F^{Image}_{proposals+}\|_2^2, \quad (8)$$

where F_{FPN} represents the feature pyramid obtained through FPN, $F^{RoI}_{proposals+}$ denotes the features of proposals containing latent novel-category objects after the RoI, and $\mathcal{L}_{\text{kd}}$ represents the distillation loss. This process effectively integrates novel-category knowledge into the detector.

3.4. Re-Weighted RPN (R-RPN) during Testing

To mitigate premature filtering of novel-category proposals during RPN post-processing at test time due to low confidence scores, we adjust proposal confidence before post-processing to boost proposals containing latent novel-category objects and improve their recall. Specifically, we first process the NMS results in the RPN. **We then apply the method in Sec. 3.3 to classify post-NMS proposals as foreground objects**, and combine the resulting confidence with the original RPN confidence as

$$S_{R\text{-}RPN} = \alpha \cdot S_{RPN} + (1 - \alpha) \cdot S_{K\text{-}FPN}, \quad (9)$$

where α is the confidence weight that controls the weighted ratio between the RPN confidence and K-FPN confidence, S_{RPN} represents the foreground object confidence from the original RPN, $S_{K\text{-}FPN}$ represents the foreground object confidence from K-FPN, and $S_{R\text{-}RPN}$ denotes the confidence of the proposals after the weighted fusion.

Ultimately, we replace the original RPN confidence with the weighted confidence for subsequent processing steps. The subsequent process follows the standard RPN post-processing pipeline: the fused scores are sorted in descending order, and the top 1,000 proposals are retained and passed to the RoI head for classification. This ensures that the proposals passed to the RoI head effectively encompass latent novel-category objects, thereby enhancing the recall of novel-category objects.

3.5. Framework and Details

Training stage. To reduce training time and computational overhead, we first adopt the RPN trained on base categories (base-RPN) in CLIPSelf to generate proposals, as base-RPN has been shown in ViLD [7] to detect the vast majority of foreground objects. Subsequently, following the approach in Sec. 3.3, we use 30 foreground and 30 background text prompts—an already diverse and balanced set that covers most semantic types in open-world scenarios, with little benefit from further expansion—to select the top 100 candidate boxes whose similarity to foreground prompts exceeds that to background prompts. Choosing 100 candidates strikes an effective balance between coverage and efficiency, capturing potential novel objects while avoiding unnecessary computational and storage overhead. Next, we rescale these selected proposals to the original image resolution, store them, crop the corresponding image regions, and feed the crops into CLIP to extract and save their features for subsequent knowledge self-distillation. During training, if a proposal exhibits high overlap with any cached candidate, we replace it with the cached candidate before feeding it into the RoI head, while the remaining proposals are processed following the standard detection pipeline.

In the self-distillation branch, we use the stored candidates and their features for distillation supervision. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg-RPN}} + \mathcal{L}_{\text{cls-RPN}} + \mathcal{L}_{\text{reg-RoI}} + \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{kd}}, \quad (10)$$

where $\mathcal{L}_{\text{cls-RPN}}$, $\mathcal{L}_{\text{reg-RPN}}$, and $\mathcal{L}_{\text{reg-RoI}}$ denote the standard classification and regression losses for detection, and $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{kd}}$ represent the contrastive loss and knowledge self-distillation loss for the RoI head, respectively. The $\mathcal{L}_{\text{kd}}$ employs the $\mathcal{L}_2$ loss. The weight of the self-distillation loss is set to 1 because we consider novel-category distillation to be as important as base-category alignment, while the remaining loss weights follow the F-ViT.

We adopt the optimized CLIP ViT-B/16 and ViT-L/14 from CLIPSelf [35] and its improved version DeCLIP [32] as our backbones, and keep them (both the image and text encoders) frozen. During training, we only train the detection modules (FPN, RPN and RoI). Simultaneously, we interpolate the feature maps from layers $[3, 5, 7, 11]$ of ViT-B/16 with relative scales $\left[\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\right]$ to the input image size. For ViT-L/14, we interpolate the feature maps from layers $[6, 10, 14, 23]$ with relative scales $\left[\frac{1}{3.5}, \frac{1}{7}, \frac{1}{14}, \frac{1}{28}\right]$ to the input image size. In Fig. 3, the weights W is set to 0.3. We train the model for 5 epochs on the OV-COCO and for 50 epochs on the OV-LVIS. We use 16 NVIDIA 3090 GPUs, with a batch size of 10 per GPU, and we use the AdamW optimizer with a learning rate of 10^{-4} and a weight decay of 0.1.

Testing stage. During inference, we use text prompts for both *base* and *novel* categories. R-RPN recalibrates proposal scores with $\alpha = 0.5$ in Eq. 9. Object categories are then predicted via cosine similarity between proposal features and all text prompts. Details are provided in the **supplementary material**.

4. Experiments

4.1. Datasets and Evaluation Metrics

Our method is evaluated on the standard OVD benchmark, LVIS [8]. LVIS comprises 100K images and 1,203 categories. The categories are divided into three groups based on the number of training images: 'frequent', 'common', and 'rare'. Following ViLD [7], we treat 337 'rare' categories as *novel categories* and train exclusively on the *base categories* (405 'frequent' and 461 'common' categories). This benchmark is referred to as OV-LVIS. We follow previous work in reporting the average precision for OV-LVIS. we report the average precision for 'frequent', 'common', and 'rare' categories, denoted as AP_f, AP_c, and AP_r respectively. The symbol AP represents the average precision across all categories. The open-vocabulary COCO (OV-COCO) benchmark, proposed by OVRCNN [39], divides the 65 categories in MS COCO into 48 base and 17 novel categories. Their accuracies are represented by AP^{50}_{base} and AP^{50}_{novel}, while AP^{50} represents the average precision.

4.2. Comparison Methods

For a fair comparison, we select several mainstream VLM-based OVD methods for testing, evaluation, and comparison on the OVD benchmarks. Specifically, we include the transfer learning approaches, *i.e.*, OWL-ViT [20], F-VLM [15], CLIPSelf [35], DST-Det [38], LBP [17], OV-DQUO [31] and DeCLIP [32], and several knowledge distillation methods, *i.e.*, OADP [33], RKDWTF [1], DK-DETR [18], RegionCLIP [43], pseudo-labeling method MM-OVOD [12], Detic [45], SAS-Det [42], and region-aware training method RO-ViT [13], CORA [36], YOLO-World [2] and YOLOE [30], which retrain a network from scratch using large-scale datasets. We adopt the optimized CLIP ViT-B/16 and ViT-L/14 from CLIPSelf [35] and its improved version DeCLIP [32] as our backbones and apply our approach on top of them, using the F-ViT proposed in CLIPSelf as the baseline detection framework.

Table 1. Comparison with SOTA methods on OV-LVIS (%).

Method	Backbone	Training Data	AP_f	AP_c	AP_r	AP
RegionCLIP	RN50*	CC3M	34.0	27.4	17.1	28.2
	RN50x4*		36.9	32.1	22.0	32.3
Detic	RN50*	LVIS-base + IN-L	-	-	24.9	32.4
OWL-ViT	ViT-B/16	O365 + VG	-	-	20.6	27.2
	ViT-L/14		-	-	31.2	34.6
RKDWTF	RN50* Base	LVIS-base + IN-L	26.4	19.4	12.2	20.9
	RN50* RKDPIS		25.5	20.9	17.3	22.1
	RN50* WTF		26.7	21.4	17.1	22.8
	RN50* WTF8x		29.1	25.0	21.1	25.9
OADP	ViT-B/32	LVIS-all	32.0	28.4	21.9	28.7
DK-DETR	RN50	LVIS-all	40.2	32.0	22.2	33.5
CORA	RN50x4	LVIS-base	-	-	22.2	-
CORA+		LVIS-base + IN-21K	-	-	28.1	-
RO-ViT	ViT-B/16	ALIGN	-	-	28.0	30.2
	ViT-L/16		-	-	32.1	34.0
	ViT-H/16		-	-	34.1	35.1
YOLO-World	YOLOv8-S*	O365 + GoldG	26.3	16.3	13.5	19.7
	YOLOv8-M*		32.7	22.5	19.9	26.0
	YOLOv8-L*		35.4	24.9	22.9	28.7
YOLOE	YOLOv11-S*		29.3	26.8	21.4	27.5
	YOLOv11-M*		34.5	32.5	26.9	33.0
	YOLOv11-L*		36.5	35.0	29.1	35.2
MM-OVOD	RN50*	LVIS-base	-	-	19.3	30.3
			-	-	18.3	29.2
			-	-	19.3	30.6
		LVIS-base + IN-L	-	-	25.8	32.7
			-	-	23.8	31.3
			-	-	27.3	33.1
F-VLM	RN50	LVIS-base	-	-	18.6	24.2
	RN50x4		-	-	26.3	28.5
	RN50x16		-	-	30.4	32.1
	RN50x64		-	-	32.8	34.9
DST-Det	ViT-B/16		-	-	26.2	-
SAS-Det	RN50-C4		31.6	26.1	20.9	27.4
	RN50x4-C4		36.8	32.4	29.1	33.5
LBP	-		32.4	28.8	22.2	29.1
OV-DQUO	ViT-B/16		23.8	27.7	29.4	26.5
	ViT-L/14		28.5	36.0	39.5	33.7
CLIPSelf + F-ViT	ViT-B/16		29.1	21.8	25.3	25.2
	ViT-L/14		35.6	34.6	34.9	35.1
CLIPSelf† + F-ViT	ViT-B/16		29.3	21.8	25.4	25.4
	ViT-L/14		35.7	34.8	35.0	35.2
CLIPSelf + **NoOVD (Ours)**	ViT-B/16		30.7	22.6	28.3 (+2.9)	26.7
	ViT-L/14		37.2	35.9	37.8 (+2.8)	36.7
DeCLIP + F-ViT	ViT-B/16		29.8	22.4	26.8	26.0
	ViT-L/14		36.5	35.2	37.2	36.0
DeCLIP† + F-ViT	ViT-B/16		29.6	22.5	26.6	26.0
	ViT-L/14		36.8	35.3	36.9	36.2
DeCLIP + **NoOVD (Ours)**	ViT-B/16		31.2	23.7	29.2 (+2.6)	27.6
	ViT-L/14		38.0	36.9	39.2 (+2.3)	37.7

Notes: IN-L denotes the inclusion of images corresponding to the 997 categories shared between ImageNet-21k-P [27] and LVIS, '*' indicates that the backbone is not initialized with CLIP, and '†' represents the results of our reproduction of CLIPSelf and DeCLIP. CC3M [29], GoldG [11], VG [14], and ALIGN [10] are all publicly available datasets.

4.3. Results on OV-LVIS

Table 1 shows the results of all comparative methods and ours on OV-LVIS.

We first observe that NoOVD + DeCLIP ViT-L/14 achieves the best performance among all competitors. Specifically, with ViT-L/14 (304.43M), NoOVD surpasses F-ViT built on the same backbones (CLIPSelf and DeCLIP) by 2.8% and 1.5% on 'rare' and overall categories, and by 2.3% and 1.5% respectively. For the smaller network ViT-B/16 (86.26M), NoOVD also outperforms F-ViT using the same backbones, improving 'rare' and overall categories by 2.9% and 1.3%, and by 2.6% and 1.6% respectively. These results validate the significant advantage of using proposals containing latent novel-category objects for knowledge self-distillation during training, particularly in

Table 2. Comparison with SOTA methods on OV-COCO (%).

Method	Backbone	Training Data	AP^{50}_{base}	AP^{50}_{novel}	AP^{50}
RegionCLIP	RN50*	CC3M	57.1	31.4	50.4
	RN50x4*		61.6	39.3	55.7
Detic	RN50*	LVIS-base + IN-L	47.1	27.8	45.0
OWL-ViT	ViT-B/16	O365 + VG	-	-	49.2
RKDWTF	RN50* Base	LVIS-base + IN-L	53.2	1.7	39.6
	RN50* RKDPIS		52.8	31.5	47.2
	RN50* WTF		54.0	36.6	49.4
	RN50* WTF8x		56.6	36.9	51.5
F-VLM	RN50	LVIS-base	-	28.0	39.6
OADP	ViT-B/32	LVIS-all	53.3	30.0	47.2
DK-DETR	RN50	LVIS-all	61.1	32.3	-
CORA	RN50	LVIS-base	35.5	35.1	35.4
	RN50x4		44.5	41.7	43.8
CORA+	RN50x4	LVIS-base + COCO Cap.	60.9	43.1	56.2
RO-ViT	ViT-B/16	ALIGN	-	30.2	41.5
	ViT-L/16		-	33.0	47.7
DST-Det	ViT-B/16	LVIS-base	59.6	41.3	54.8
	ViT-L/14		61.9	46.7	58.0
SAS-Det	RN50-C4		58.5	37.4	53.0
LBP	-		60.8	35.9	54.3
OV-DQUO	ViT-B/16		42.6	39.4	41.7
	ViT-L/14		48.3	45.3	47.5
CLIPSelf + F-ViT	ViT-B/16		54.9	37.6	50.4
	ViT-L/14		64.1	44.3	59.0
CLIPSelf + F-ViT†	ViT-B/16		55.8	35.8	50.6
	ViT-L/14		64.8	42.9	59.1
CLIPSelf + **NoOVD (Ours)**	ViT-B/16		56.5	37.9 (+2.1)	51.6
	ViT-L/14		64.7	45.4 (+2.5)	59.7
DeCLIP + F-ViT	ViT-B/16		57.8	41.1	53.5
	ViT-L/14		65.2	46.2	60.3
DeCLIP + F-ViT†	ViT-B/16		58.1	40.3	53.4
	ViT-L/14		65.5	45.1	60.2
DeCLIP + **NoOVD (Ours)**	ViT-B/16		58.3	41.9 (+1.6)	54.0
	ViT-L/14		65.8	47.5 (+2.4)	61.0

enhancing the generalization of novel categories in OVD. Notably, NoOVD also shows improvements in base category accuracy compared to F-ViT. We attribute this to the fact that learning novel category knowledge simultaneously strengthens the model's ability to differentiate between categories. Moreover, although methods such as RO-ViT leverage larger-scale pretraining and achieve strong performance even with smaller backbones, NoOVD delivers superior results on novel categories and larger backbones—which are the primary focus of OVD.

4.4. Results on OV-COCO

Table 2 presents the results of all comparative methods and ours on OV-COCO. We first observe that the proposed NoOVD + DeCLIP ViT-L/14 likewise achieves the best performance among all competitors. Specifically, when using ViT-L/14 with CLIPSelf and DeCLIP, NoOVD surpasses F-ViT by -0.1%, 2.5%, and 0.6% on the base, novel, and overall categories, and by 0.3%, 2.4%, and 0.8%, respectively. Furthermore, compared to F-ViT built on the CLIPSelf and DeCLIP ViT-B/16 backbones, NoOVD improves the same three groups by 0.7%, 2.1%, and 1.0%, and by 0.2%, 1.6%, and 0.6%, respectively. We also observe that the performance gains of NoOVD on OV-COCO are smaller than those on OV-LVIS. It is important to note that OV-LVIS is constructed by extending the annotations of OV-COCO, providing richer and more comprehensive labels and covering broader and larger category sets. During training, unlabeled objects in OV-COCO are not treated as background, instead, the model regards them as foreground and distills knowledge from them. However, during testing, although the model successfully detects these objects, they are counted as false positives due to missing annotations, thereby suppressing the overall accuracy. Hence, the smaller performance gains on OV-COCO primarily stem from its incomplete annotations rather than limitations of NoOVD. Therefore, for evaluating OVD, we consider OV-LVIS to be more stable and reliable than OV-COCO.

Table 3. Cross-dataset main results on Objects365 (%).

Method	Backbone	Training Data	AP_r	AP	AP^{50}
Detic	RN50*	LVIS-all	9.5	13.9	19.7
		LVIS-all + IN-L	12.4	15.6	22.2
MM-OVOD		LVIS-all	10.1	14.8	21.0
		LVIS-all + IN-L	13.1	16.6	23.1
CLIPSelf + F-ViT	ViT-B/16	LVIS-base	16.8	19.0	32.3
	ViT-L/14		21.7	23.7	39.2
CLIPSelf + **NoOVD (Ours)**	ViT-B/16		18.3	19.6	33.0
	ViT-L/14		22.8	24.6	40.2
DeCLIP + F-ViT	ViT-B/16		17.6	20.2	33.1
	ViT-L/14		22.3	24.5	39.8
DeCLIP + **NoOVD (Ours)**	ViT-B/16		19.0	21.1	34.0
	ViT-L/14		23.3	25.3	41.1

4.5. Cross-Dataset Transfer Results

Table 3 presents the cross-dataset transfer results from OV-LVIS to Objects365. We compare NoOVD with models from Detic, MM-OVOD, and CLIPSelf / DeCLIP + F-ViT, reporting the bounding box AP metric as the standard on Objects365. In all cases, the Detic and MM-OVOD models are trained on LVIS-all, with IN-L models using ImageNet-21k-P as additional weak supervision, while CLIPSelf / DeCLIP + F-ViT and NoOVD are tested with models trained on the base categories of OV-LVIS. The trained open-vocabulary detectors are evaluated on the Objects365 validation set. Following MM-OVOD, we define the bottom third of the categories in Objects365, based on frequency, as 'rare' categories.

Across all settings, CLIPSelf / DeCLIP + F-ViT already surpasses MM-OVOD and Detic. When using ViT-B/16 as the backbone, NoOVD improves over CLIPSelf + F-ViT by 1.5% in AP_r and 0.7% in AP^{50}, and over DeCLIP + F-ViT by 1.4% and 0.9%, respectively. With ViT-L/14, NoOVD outperforms CLIPSelf + F-ViT by 1.1% in AP_r and 1.0% in AP^{50}, and surpasses DeCLIP + F-ViT by 1.0% and 1.3%, respectively. These results demonstrate that NoOVD, by identifying latent novel-category objects and training the model to adapt to a broader set of categories, significantly enhances the model's transferability.

4.6. Ablation Study

Ablation studies use CLIPSelf ViT-B/16 as the backbone.

Component analysis. As shown in Table 4, We conduct comprehensive ablation experiments on K-FPN and R-RPN using F-ViT as the baseline framework. In the experiments, 'CLIP-top' refers to directly using the top-level feature map output by the frozen CLIP ($\{P4\}$) in Fig. 3 to replace K-FPN for feature extraction. The results show that using the simple CLIP top-layer feature can improve the performance, which verifies the basic idea of NoOVD. K-FPN improves novel category detection more significantly by 2.1%, demonstrating that K-FPN, by constructing multi-scale feature maps, is more effective in discovering latent novel-category objects and facilitating knowledge self-distillation

Table 4. Effects of each module in NoOVD on the OV-LVIS (%).

Baseline	CLIP-top	K-FPN	R-RPN	AP_r	AP
✓				25.4	25.4
✓	✓			26.4	26.1
✓	✓	✓		27.5	26.4
✓			✓	26.7	25.9
✓	✓		✓	27.8	26.4
✓	✓	✓	✓	28.3	26.7

in the network. Meanwhile, the use of R-RPN directly enhances the baseline method by 1.3% in novel category detection, indicating that improving the recall rate of novel-category objects in the RPN stage during testing is essential. This helps prevent novel-category objects from being filtered out during detection, which would otherwise affect subsequent classification. Ultimately, the combination of K-FPN and R-RPN yields the best performance.

Table 5. Sentitivity analysis of hyperparameter W (%).

W	0.2	0.3	0.4	0.5	0.7	0.8
AP	26.2	26.7	26.3	24.8	22.5	19.1

Feature fusion weights of K-FPN. We conduct ablation studies on the fusion weights of different K-FPN feature levels. As shown in Table 5, the best performance occurs at $W = 0.3$. When W is too small, high-level features dominate excessively and suppress low-level details, leading to performance degradation; when W is too large, high-level semantic information becomes insufficient, and accuracy similarly drops. This confirms that high-level features provide strong semantic abstraction, whereas low-level features contribute essential structural and textural details. This observation is consistent with prior understanding: high-level features offer stronger semantic abstraction, whereas low-level features preserve richer local details, making them naturally complementary. An appropriate fusion ratio strikes an effective balance between semantics and detail, significantly improving overall performance.

The fusion weight of R-RPN. We conduct ablation studies on the fusion weight α in Eq. 9. As shown in Table 6, $\alpha = 0.5$ yields the best performance, which also aligns with our intuition. We argue that the K-FPN and RPN scores play equally important roles in R-RPN, and assigning them equal weights establishes a balanced and effective fusion mechanism. Under this strategy, base-category objects retain high fused scores because both K-FPN and RPN provide strong confidence, ensuring they are reliably forwarded to the RoI stage. For latent novel-category objects, although their RPN scores are often low, the strong semantic cues from K-FPN—derived from VLM knowledge—elevate their fused scores, substantially increasing their likelihood of being preserved. Meanwhile, proposals classified as background by both branches maintain low fused scores and are naturally filtered out. Overall, $\alpha = 0.5$ strikes an effective balance between base and novel categories, preserving reliable detection of known classes while significantly improving the recall of latent novel objects.

Choice of distillation loss function. As shown in Table 7, we conduct an ablation study on the loss function $\mathcal{L}_{\text{kd}}$ used for knowledge self-distillation. We explore several loss functions, including $\mathcal{L}_1$, $\mathcal{L}_2$, $\text{Smooth}\mathcal{L}_1$, and $\mathcal{L}_{\text{cons}}$ (cosine similarity loss). The experimental results indicate that $\mathcal{L}_2$ yields the best performance, while $\mathcal{L}_1$ and $\text{Smooth}\mathcal{L}_1$ also produce similar results. However, using $\mathcal{L}_{\text{cons}}$ for alignment performs relatively poorly. We argue that the $\mathcal{L}_2$, being more sensitive to large deviations, can more effectively minimize the distance between RoI features and CLIP representations, thereby promoting precise alignment within the semantic space. In contrast, $\mathcal{L}_1$ treats all errors uniformly, while $\text{Smooth}\mathcal{L}_1$ is overly gentle in regions of small discrepancy, potentially hindering the convergence of cross-modal differences. Moreover, training with $\mathcal{L}_{\text{cons}}$ may fail to sufficiently align the feature representations, leaving a significant information gap between them. This misalignment can impair the acquisition of novel category knowledge and ultimately compromise detection performance.

Table 6. Sentitivity analysis of hyperparameter α (%).

α	0.9	0.7	0.5	0.3	0.1
AP	25.5	25.8	26.7	26.2	25.9

Table 7. Results of different knowledge distillation losses (%).

$\mathcal{L}_{\text{kd}}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\text{Smooth}\mathcal{L}_1$	$\mathcal{L}_{\text{cons}}$
AP_r (%)	27.8	28.3	28.0	17.3

Extra computation cost. NoOVD's extra cost compared to F-ViT comes from proposal cropping and feature extraction, both carried out offline before training. We use an RPN trained on base categories (replaceable by a stronger variant) to generate proposals and extract all proposal features in one pass using CLIP ViT-B/16 on 8×3090 GPUs (53 minutes in total). Because each image's novel-category proposals are fixed, training only loads cached features for distillation, requiring no repeated cropping or forward passes. Other than this offline step, the overall training time is effectively the same as F-ViT.

Additional computation cost analysis and visualizations are included in the supplementary materials.

5. Conclusion

In this work, we have thoroughly examined the issues present in current OVD frameworks, particularly the category gap between training and testing phases. To this end, we propose an novel framework, NoOVD, which leverages the knowledge from frozen VLMs to uncover latent novel-category objects, while integrating knowledge self-distillation to prevent the forced alignment of novel categories with the background. This approach allows the model to learn novel-category knowledge based on the latent novel-category objects. During testing, we have enhanced the recall rate of novel-category objects by adjusting the confidence of proposals. Extensive experiments validate the effectiveness of NoOVD, demonstrating superior performance. Our research introduces a new schema for OVD.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U2574216, 62402490, and 62506146; in part by the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010101; in part by the Jiangxi Provincial Natural Science Foundation under Grant 20252BAC200196.

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, 6

[2] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2, 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 248–255, 2009. 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] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[6] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 3

[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. 1, 2, 3, 5, 6

[8] 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, 6

[9] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2462–2470, 2024. 2

[10] 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, 2021. 1, 2, 3, 6

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 6

[12] 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. 3, 6

[13] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11144–11154, 2023. 2, 6

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[15] 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. 1, 2, 3, 6

[16] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[17] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. 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*, pages 16678–16687, 2024. 2, 6

[18] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510, 2023. 1, 2, 3, 6

[19] 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, 2014. 1, 2

[20] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*, 2, 2022. 1, 6

[21] Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788, 2024. 1

[22] Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. Dovtrack: Data-efficient open-vocabulary track-

ing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[23] Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. Covtrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10054–10063, 2025. 3

[24] 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, 2021. 1, 2, 3

[25] 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. 3

[26] 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*, 28:91–99, 2015. 1

[27] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[28] 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. 2

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[30] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*, 2025. 2, 6

[31] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7762–7770, 2025. 1, 2, 6

[32] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 1, 2, 3, 6

[33] 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, 6

[34] 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.

[35] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 1, 2, 3, 4, 6

[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, 6

[37] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[38] Shilin Xu, Xiangtai Li, Size Wu, Wenwei Zhang, Yunhai Tong, and Chen Change Loy. Dst-det: Open-vocabulary object detection via dynamic self-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2, 6

[39] 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. 6

[40] Yupeng Zhang, Shuqi Zheng, Ruize Han, Yuzhong Feng, Junhui Hou, Linqi Song, Wei Feng, and Liang Wan. Rethinking the one-shot object detection: Cross-domain object search. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9573–9581, 2024. 3

[41] Yupeng Zhang, Ruize Han, Fangnan Zhou, Song Wang, Wei Feng, and Liang Wan. Odov: Towards open-domain open-vocabulary object detection. *arXiv preprint arXiv:2508.01253*, 2025. 3

[42] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Yumin Suh, Manmohan Chandraker, Dimitris N Metaxas, et al. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13938–13947, 2024. 2, 6

[43] 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. 6

[44] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3

[45] 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, 2022. 2, 6