Explicit Semantic–Spatial Alignment for Open-Vocabulary Object Detection

Pengyang Su[1], Weihong Ren[1⋆], Shuhuan Han[1], Qian Dong[1], Haoran Xu[1], Xi Ai[1], Zijian Wang[1], Zhiyong Wang[1], and Honghai Liu[1,2]

[1] Harbin Institute of Technology, Shenzhen, China
25s053010@stu.hit.edu.cn, weihongren@hit.edu.cn
[2] Southeast University, Nanjing, China

Abstract. Open-Vocabulary object Detection (OVD) typically relies on Vision–Language Models (VLMs) to associate visual regions with arbitrary textual concepts. To compensate for the limited localization capability of VLMs, existing approaches often incorporate large pre-trained visual features that inherently encode object-centric spatial cues. However, these features are commonly introduced through direct fusion or shallow adaptation, which fails to explicitly align semantic representations with object-level spatial information, resulting in imprecise localization. In this paper, we propose a novel framework for explicit semantic–spatial alignment in open-vocabulary object detection. Rather than directly merging pre-trained object-centric visual features, our method progressively aligns semantic representations with object-level spatial cues through a staged alignment process. Specifically, we introduce a lightweight spatial adapter that spatially recalibrates auxiliary object-centric visual features to suppress background responses and emphasize salient object regions. Building upon this, we design a frequency-aware fusion mechanism that decomposes the adapted object-centric visual features into spectral components and adaptively injects object-sensitive spatial cues into semantic features at multiple levels. This targeted fusion strategy enables effective semantic alignment while preserving the original semantic embedding space. Extensive experiments on the OV-COCO and OV-LVIS benchmarks demonstrate the effectiveness of our approach against state-of-the-art methods.

Keywords: Open-Vocabulary Object Detection · Semantic–Spatial Alignment · Frequency-Aware Feature Fusion · Vision–Language Models

1 Introduction

Object detection, a fundamental perception task in computer vision, has traditionally been confined to a closed-world assumption, where models are trained and tested on a fixed set of predefined categories [28,29]. This restriction severely limits their deployment in dynamic real-world scenarios. To break these shackles, Open-Vocabulary object Detection (OVD) has emerged as a transformative

⋆ Corresponding author

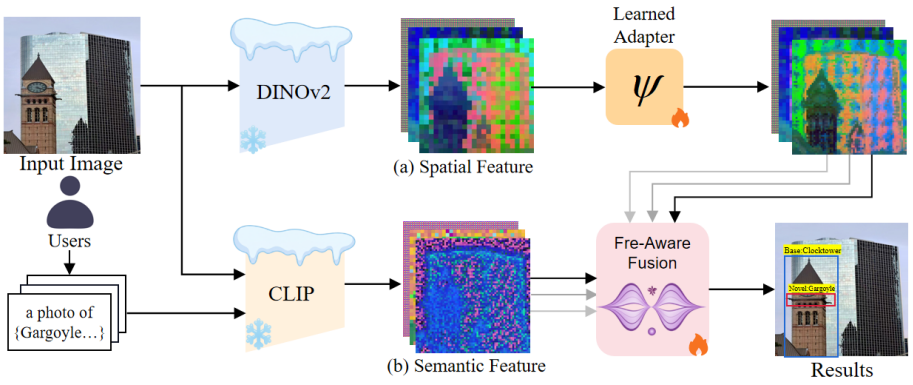

Fig. 1: The motivation to eliminate the Semantic-Spatial Discrepancy. The feature maps of the spatial alignment feature (a) highlight precise spatial structure and sharp object boundaries, while the feature maps of the semantic feature (b) exhibit strong semantic activation with coarse localization. How to effectively integrate them together is crucial for OVD.

paradigm [8, 39]. Drawing on the impressive zero-shot capabilities of large-scale Vision-Language Models (VLMs) like CLIP [27] and ALIGN [13], recent OVD initiatives project images and texts into a unified embedding space via internet-scale contrastive learning. This empowers detectors to recognize arbitrary novel objects driven by free-form natural language queries.

However, adapting these image-level VLMs for dense, region-level prediction exposes a fundamental paradox: the Semantic-Spatial Discrepancy. Pretrained through image-text contrastive objectives, VLMs inherently act as a "bag of concepts" [20]. They excel at aggregating global semantics for cross modal alignment but systematically suppress fine-grained structural details to achieve spatial invariance [4]. As illustrated in Fig. 1(b), the feature activations from CLIP provide strong semantic identity cues but suffer from coarse localization and diffuse boundaries, inevitably leading to inaccurate bounding box regression and susceptibility to environmental artifacts. In contrast, feature maps of DINOv2 preserve fine-grained spatial structure and exhibit sharper object boundaries, offering object centric spatial cues that are more amenable to precise localization. These observations highlight the need to reduce the Semantic–Spatial Discrepancy in OVD by explicitly aligning open-vocabulary semantics with object level spatial information.

To mitigate this localization degradation, prior works typically resort to knowledge distillation [8, 44], pseudo-labeling via self-training [21, 39], or prompt tuning [5, 36]. While these methods force the model to attend to local regions through external supervision, they may treat the symptoms rather than the root cause: the intrinsic spectral deficiency of the CLIP-like representations. It is well established that high-frequency signal components capture the sharp edges and boundary discontinuities essential for spatial delineation [37]. Stan-

dard contrastive pre-training effectively acts as a low-pass filter, attenuating high-frequency spatial priors while retaining global semantic concepts. Conversely, Visual Foundation Models (VFMs) such as DINOv2 [26], trained with self-supervised objectives at the patch level, learn spatial-aware representations that better preserve local structure. As shown in Fig. 1(a), feature maps of DINOv2 exhibit striking spatial precision. Therefore, overcoming the OVD bottleneck requires explicitly recovering the missing spatial spectra by complementing semantic representations with structure-aware cues.

To bridge this gap, in this paper, we propose ESSA-OVD, a novel framework for Explicit Semantic–Spatial Alignment for Open-Vocabulary object Detection. Our core philosophy is to synergize the cross-modal semantic flexibility of CLIP with the fine-grained spatial integrity of DINOv2 via a dual-stream architecture. Specifically, we introduce a Spatial Adapter to project the spatial-rich features into a manifold compatible with the CLIP embedding space, bridging the inherent domain gap. Following this alignment, a Frequency-Aware Fusion (FAF) mechanism selectively isolates high-frequency spatial cues and explicitly injects them into the intermediate layers of the semantic stream. This effectively sharpens the detector's boundary perception at the feature level. Furthermore, to prevent "semantic drift" where multi-modal fusion might corrupt the pre-trained zero-shot classification space, we enforce a strict Pure Dense Strategy. This ensures that the final dense feature map used for text matching remains pristine, effectively decoupling precise localization from robust semantic recognition. Our main contributions are summarized as follows:

- We characterize the semantic–spatial discrepancy in OVD and address the resulting localization bottleneck by explicitly injecting DINOv2 spatial priors into CLIP features, while introducing only **3M** additional trainable parameters beyond the detector head..
- We propose the ESSA-OVD framework, introducing a Spatial Adapter and a Frequency-Aware Fusion (FAF) module to explicitly align and inject multi-layers high-frequency spatial details into the semantic stream without corrupting its language alignment.
- We introduce a Pure Dense strategy that explicitly decouples feature routing for localization and classification, ensuring the zero-shot semantic coherence of CLIP is strictly preserved.
- Extensive experiments show that ESSA-OVD achieves state-of-the-art performance on **OV-COCO** and **OV-LVIS** benchmarks, and also improves cross-dataset transfer to **Objects365** dataset.

2 Related Work

Open-Vocabulary Object Detection. The goal of OVD is to detect novel objects beyond the base categories seen during training. Existing approaches typically bridge the gap between closed-set detectors and open-world VLMs via two main paradigms: feature distillation and pseudo-labeling. Distillation-based methods, such as ViLD [8], align the detector's region embeddings with the im-

age embeddings of CLIP, by transferring semantic knowledge from teacher to student. Building on this, CCKT [41] recently introduced a cyclic knowledge transfer framework. It employs a regional contrastive loss to refine query representations without requiring extra image-text pairs, effectively enhancing the detector's sensitivity to novel concepts. Conversely, pseudo-labeling approaches like OVR-CNN [39] generate supervision signals for unannotated objects. Addressing the confidence bias in these methods, OV-DQUO [32] proposes a denoising text query training strategy combined with wildcard matching. This allows the model to leverage unknown objects in the open world as supervision, significantly reducing the interference from background noise. While these state-of-the-art methods optimize semantic alignment or training strategies, they largely treat the pre-trained VLM features as given. They overlook the intrinsic spectral deficiency of these features, specifically the lack of high-frequency spatial information, which limits the upper bound of localization precision.

Localization via Vision-Language Models. VLMs like CLIP [27] and ALIGN [13] are trained to match global image feature with text description, resulting in "bag-of-concepts" representations [20] that are invariant to spatial structure. To adapt these models for dense prediction, RegionCLIP [44] proposes region-based pre-training to explicitly align local image patches with text concepts. Other works, such as DetPro [5] and CORA [36], utilize prompt learning to adapt the VLM's attention mechanisms to downstream detection tasks. Although these methods attempt to mitigate the localization gap through external constraints or adaptation, they do not alter the fundamental frequency characteristics of the underlying features. Our work argues that without explicitly recovering the missing high-frequency structural signals, the "drifting box" problem cannot be fundamentally resolved.

Visual Foundation Models in OVD. Self-supervised Visual Foundation Models (VFMs), particularly DINO [1] and DINOv2 [26], have emerged as powerful tools for spatial tasks. Trained with patch-level discriminative objectives, DINOv2 exhibits strong object-centric priors compared to contrastive VLMs. Recent works have started to incorporate VFMs into detection pipelines, e.g., leveraging SAM [17] or DINO [1] cues to refine masks or proposals [10, 14]. However, these integrations are mostly designed for refinement and often employ coarse fusion (e.g., concatenation or late fusion), leaving it under-explored how to inject spatial cues while maintaining CLIP's text–image alignment required for zero-shot classification. In contrast, we fuse VFM cues through a spectral lens, selectively injecting spatial priors while explicitly preserving the semantic integrity of the VLM.

3 Methodology

In this section, we propose ESSA-OVD, a framework designed to resolve the spectral imbalance in open-vocabulary detection. Fig. 2 provides an overview of our architecture. We first outline the problem setup and baseline detector in Sec. 3.1. Then, we introduce our dual stream framework in Sec. 3.2, followed

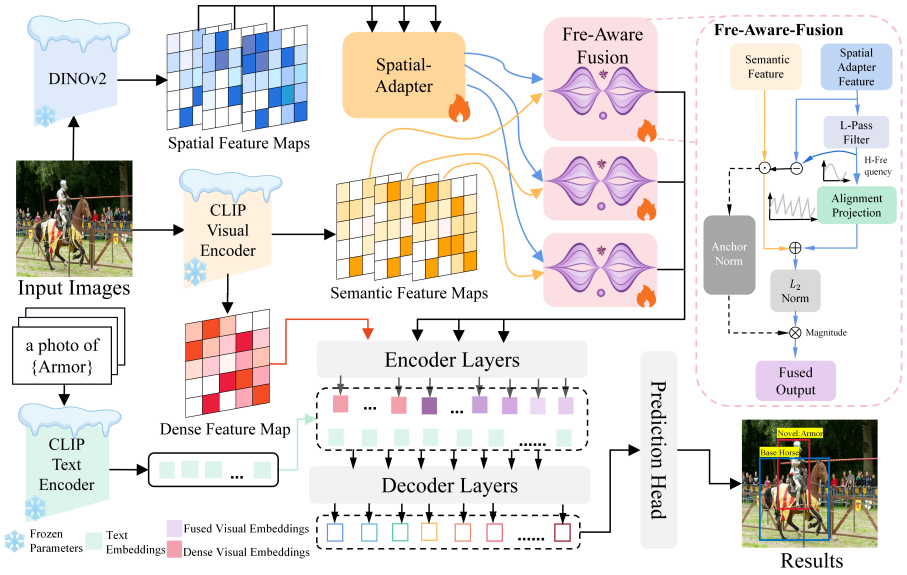

Fig. 2: The overall framework of ESSA-OVD. A dual-stream backbone with a frozen CLIP [27] semantic stream and a frozen DINOv2 [26] spatial stream. Spatial Adapter aligns heterogeneous feature manifolds, and the Frequency-Aware Fusion (FAF) module injects high-frequency structural priors into the semantic stream. Pure Dense Strategy applies fusion only to the encoder-oriented multi-level layer features, while keeping the final classification-oriented dense feature (F_{cls}) pure to preserve semantic alignment for robust open-vocabulary generalization. For clarity, IsoNorm is shown as L_2 normalization and anchor-norm rescaling.

by detailed formulations of the Spatial Adapter, Frequency-Aware Fusion, and Pure Dense Strategy in subsequent sections.

-

3.1 Preliminaries

Problem Definition. Following the standard OVD protocol [8, 32], we divide object categories into base classes $\mathcal{C}_B$ and novel classes $\mathcal{C}_N$, where $\mathcal{C}_B \cap \mathcal{C}_N = \emptyset$. The detector is trained on the base dataset $\mathcal{D}_{train}$ containing only annotations for $\mathcal{C}_B$, but must generalize to detect objects from $\mathcal{C}_B \cup \mathcal{C}_N$ during inference. We leverage a pre-trained CLIP model to generate text embeddings $\mathcal{T} \in \mathbb{R}^{K \times D}$ as the classifier weights.

Revisiting Deformable DETR. We adopt Deformable DETR [46] as our baseline. It utilizes a Transformer encoder-decoder architecture to transform image features into a set of object predictions. Let $y = \{(c_i, b_i)\}$ be the ground truth set and $\hat{y} = \{(\hat{p}_i, \hat{b}_i)\}$ be the set of N predictions. The training objective minimizes the Hungarian matching cost to find an optimal bipartite matching

σ:

$$\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) = -\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{box}(b_i, \hat{b}_{\sigma(i)}), \tag{1}$$

where $\mathcal{L}_{box}$ combines L1 loss and GIoU [40] loss. This end-to-end paradigm allows us to focus on optimizing the feature representation for both localization and recognition.

3.2 Overview of ESSA-OVD

Existing methods typically rely on a single frozen CLIP backbone, which creates a *Semantic-Spatial Discrepancy*: Semantic features act as low-pass filters that blur structural details required for precise regression. To bridge this gap, ESSA-OVD introduces a Dual Stream paradigm. We employ a frozen CLIP to preserve open-world semantic knowledge (Semantic Stream) and a parallel frozen DINOv2 to supply high-frequency structural priors (Spatial Stream). These heterogeneous streams are synergized via an Explicit Semantic-Spatial Alignment mechanism: raw multi-layers spatial features are first rectified by a Spatial Adapter, then spectrally filtered by a Frequency-Aware Fusion (FAF) module, and finally routed via a Pure Dense Strategy to decouple localization from classification.

3.3 Spatial Adapter

A naive element-wise addition of $\mathcal{F}^{sem}$ and $\mathcal{F}^{spa}$ is theoretically suboptimal due to the significant domain gap. Spatial features are optimized for instance discrimination, residing in a latent manifold distinct from the language-aligned space of semantic feature. Furthermore, there exists a discrepancy in channel dimensions (e.g., $C_{spa} = 1024$ vs. $C_{sem} = 768$). Therefore, before any interaction, the spatial features must be rectified to be statistically compatible with the semantic stream.

We propose a lightweight **Spatial Adapter**, denoted as $\mathcal{A}(\cdot)$, to project the raw spatial features into the semantic feature space. For a feature map $F_l^{spa} \in \mathbb{R}^{C_{in} \times H_l \times W_l}$ at layer l, the adaptation process is formulated as:

$$\hat{F}_l^{spa} = \mathcal{A}(F_l^{spa}) = \sigma\left(\mathrm{GN}_G\left(\mathbf{W}_{proj} F_l^{spa}\right)\right), \tag{2}$$

where $\mathbf{W}_{proj}$ represents a 1×1 convolution that compresses the channel dimension from C_{in} to C_{out}. GN_G denotes Group Normalization with $G = 32$ groups, and σ is the ReLU activation function. The inclusion of Group Normalization is critical. Unlike Batch Normalization, GN is independent of batch size and robustly standardizes the internal covariate shift of DINOv2 features. This ensures that the injected spatial signal has a consistent distribution (zero mean and unit variance) before entering the semantic stream, preventing the auxiliary features from overwhelming the primary semantic representation due to magnitude mismatch.

3.4 Frequency-Aware Fusion

After alignment, simply fusing the entire spectrum of $\hat{F}_l^{spa}$ into the semantic stream is still problematic. DINOv2 features, while spatially rich, may contain texture patterns or global layout information that are redundant or even conflicting with CLIP's high-level semantics. We argue that the most valuable complementary information lies specifically in the high-frequency spectral band, which encodes object boundaries and edges essential for localization. Based on this insight, we propose the FAF module.

Spectral Decomposition. We explicitly separate the adapted spatial feature into low-frequency and high-frequency(proxy) components. Inspired by the unsharp masking technique in signal processing, we approximate the low-frequency component (background/layout) using a local smoothing operator. The high-frequency residual $\mathbf{H}_l$, which captures the structural details, is derived as:

$$\mathbf{H}_l = \hat{F}_l^{spa} - \mathcal{P}_{low}(\hat{F}_l^{spa}), \quad \text{where } \mathcal{P}_{low}(\mathbf{x}) = \text{AvgPool}_{k\times k}(\mathbf{x}). \tag{3}$$

Here, $\mathcal{P}_{low}$ is implemented as an Average Pooling layer with kernel size $k = 3$. By subtracting the smoothed low-frequency approximation from the original feature, $\mathbf{H}_l$ effectively acts as a high-pass filtered map, highlighting the "structural residue" such as object contours while suppressing the smooth low-frequency variations.

Aligned Injection. To integrate this high-frequency prior into the semantic stream, we employ a linear projection layer $\mathcal{W}_{align}$ to further refine the features. The injection is controlled by a learnable scalar gate α, initialized to 0 as a conservative design choice. This zero-initialization strategy is crucial, as it ensures the model starts training from the stable baseline of the original semantic features and gradually learns to incorporate spatial cues:

$$\tilde{F}_l^{sem} = F_l^{sem} + \alpha \cdot \mathcal{W}_{align}(\mathbf{H}_l). \tag{4}$$

Magnitude Preservation. A critical challenge in multi-modal fusion is *Semantic Drift*. Although CLIP measures text–image similarity by cosine similarity (direction-dependent and scale-invariant), the downstream projection/Transformer blocks are sensitive to feature *scale*. Thus, the direct residual addition in Eq. (4) may change the feature norm and shift the representation statistics, which can weaken the pre-trained alignment to text embeddings.

To reduce such *scale*-induced drift, we introduce a norm-preserving rescaling operation:

$$F_l^{fused} = \text{IsoNorm}(\tilde{F}_l^{sem}, F_l^{sem}) = \frac{\tilde{F}_l^{sem}}{\|\tilde{F}_l^{sem}\|_2 + \epsilon} \cdot \|F_l^{sem}\|_2. \tag{5}$$

IsoNorm decouples direction from scale via L_2-normalization and then restores the original feature norm, where the norm is computed along the channel dimension at each spatial location [7, 18]. This does not guarantee preserving the semantic embedding distribution or cosine neighborhoods (direction changes still matter), but it stabilizes feature scale in practice; semantic robustness is further supported by our gated injection and the Pure Dense Strategy.

3.5 Pure Dense Strategy

In Deformable DETR, the semantic stream features serve two downstream purposes with conflicting requirements. The Transformer Encoder consumes hierarchical multi-level features $\{F_l\}_{l=1}^{3}$ for localization and box regression, where spatial variation and structural details are important. Here, "multi-scale" refers to the detector-side multi-level feature interface, rather than a native feature pyramid produced by the ViT backbones. In contrast, the final dense feature map F_{cls} is used for classification via text–image similarity, which requires semantic invariance and strict alignment with the text embedding space.

Using a unified feature map for both purposes is suboptimal in OVD: contaminating the classification-oriented feature with spatial cues, even with Magnitude Preservation, can still introduce subtle semantic noise. Therefore, we propose the Pure Dense Strategy, a task-decoupled feature routing scheme over the multi-level feature interface. Formally, let Φ_{FAF} denote the fusion process described in Sec. 3.4. The routing is defined as:

$$\begin{cases} F_l^{out} = \Phi_{FAF}(F_l^{sem}, F_l^{spa}), & \forall l \in \{1,2,3\}, \\ F_{cls}^{out} = F_{cls}^{sem}. \end{cases} \tag{6}$$

That is, FAF is applied only to the encoder-oriented multi-level features ($l \in \{1,2,3\}$), while the final classification-oriented feature F_{cls} bypasses fusion. This preserves the pristine CLIP semantic manifold for classification and improves transferability to novel categories. In our experiments, this decoupling consistently improves performance and avoids the degradation caused by applying spatial fusion to F_{cls}.

3.6 Training Objectives

Following the standard Deformable DETR protocol, we employ the Hungarian algorithm for bipartite matching between the N predicted proposals and the ground-truth objects. The overall training objective $\mathcal{L}_{total}$ is formulated as a weighted sum of the classification loss and the bounding box regression losses:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}, \tag{7}$$

where λ_{cls}, λ_{L1}, and λ_{giou} are the corresponding loss weight hyperparameters.

Open-Vocabulary Classification Loss ($\mathcal{L}_{cls}$). To align the visual representations with the semantic space, the classification probability p_c is computed via a softmax-normalized cosine similarity between the predicted object embedding z_q and the text embeddings of the base categories. Here, z_q is the decoder query embedding for classification, and Pure Dense acts on the encoder-side dense feature routing (by keeping $F_{cls}^{out} = F_{cls}^{sem}$) rather than replacing decoder queries. We optimize this alignment using the Focal Loss to address the severe class imbalance typically encountered in OVD:

$$p_c = \frac{\exp(\text{sim}(z_q, t_c)/\tau)}{\sum_{j \in \mathcal{C}_B \cup \{\varnothing\}} \exp(\text{sim}(z_q, t_j)/\tau)}, \quad \mathcal{L}_{cls} = -\beta(1-p_y)^{\gamma}\log(p_y), \tag{8}$$

where τ is the temperature parameter, t_c represents the text embedding of category c, $\mathcal{C}_B$ is the set of base categories, β is the class-balance factor, γ is the focusing parameter, and $\varnothing$ denotes the background class.

Bounding Box Regression Loss ($\mathcal{L}_{L1}$ and $\mathcal{L}_{\text{giou}}$). Unlike classification, the localization branch is class-agnostic. The regression loss consists of two complementary components to ensure precise box generation:

$$\mathcal{L}_{reg} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} = \lambda_{L1}\|\mathbf{b} - \hat{\mathbf{b}}\|_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}(\mathbf{b}, \hat{\mathbf{b}}), \tag{9}$$

where $\mathbf{b}$ and $\hat{\mathbf{b}}$ denote the ground-truth and predicted bounding box coordinates (i.e., center coordinates, width, and height). Specifically, $\mathcal{L}_{L1}$ computes the absolute ℓ_1 distance between the normalized coordinates to facilitate fast convergence, while $\mathcal{L}_{\text{giou}}$ measures the Generalized Intersection over Union (GIoU) to provide scale-invariant supervision for box overlap. Benefiting from the high-frequency structural cues injected by our FAF module, the regression head receives explicit boundary priors, leading to more precise localization. During inference, we substitute the base category embeddings with the text embeddings of the target vocabulary to achieve open-vocabulary detection.

4 Experiments

4.1 Datasets and Evaluation metrics

To verify the effectiveness of our proposed framework, we conduct extensive experiments on two standard open-vocabulary detection benchmarks: **OV-COCO** and **OV-LVIS**.

OV-COCO Benchmark. Following the standard OVR-CNN protocol [39], we utilize the MS-COCO 2017 dataset [22] for evaluation. The 80 object categories are split into 48 *Base classes* (annotated) and 17 *Novel classes* (unannotated). In this setting, the model is trained exclusively on the base classes, which comprise 107,761 images and 665,387 instances. Evaluation is performed on the validation set of 4,836 images, covering both base and novel categories to assess generalization. For evaluation metrics, we report the box Mean Average Precision at IoU 0.5 (AP_{50}). The primary indicator for open-vocabulary performance is the AP on novel categories, denoted as AP_{50}^{Novel}.

OV-LVIS Benchmark. To evaluate performance in a large-scale, long-tailed scenario, we adopt the LVIS v1.0 benchmark [9], which contains **1,203 categories** grouped into *Frequent*, *Common*, and *Rare* based on image frequency. Following the standard practice in ViLD [8], we treat the Frequent and Common categories (461 and 405 classes, respectively) as *Base classes*, while reserving the 337 Rare categories as *Novel classes*. The model is trained on 100,170 images containing base categories with 1,264,883 instances. Evaluation is conducted on the standard validation set of 19,809 images across all categories. We report the box mAP averaged over IoU thresholds from 0.5 to 0.95, with the mAP of rare categories (AP_r) serving as the key metric for open-vocabulary generalization.

4.2 Implement Details

Model Specifications. For the backbone, we employ a dual-stream backbone architecture. The semantic stream is initialized with the pre-trained CLIP, while the auxiliary spatial stream utilizes the robust DINOv2 [26]. To preserve the generalization capability of the pre-trained feature spaces, both backbones are frozen during training. Only the proposed Spatial Adapter, Frequency-Aware Fusion (FAF) modules, and the detector head (Transformer encoder/decoder) are trainable. *Detector.* For **fair comparison**, we follow previous open-vocabulary detection practice [31, 32, 36] in both detector configuration and training protocol: the detector uses 6 encoder layers, 6 decoder layers, and 1000 object queries (with EMA and the same schedule), and we compute category text embeddings offline using the CLIP text encoder with an ensemble of 80 prompt templates (e.g., "a photo of a {}"), projecting them to the 256-d visual feature space via a learnable MLP. Patch-token features from both ViT backbones are organized into the detector-side four-level encoder inputs $\{0, 1, 2, \text{cls}\}$; FAF is applied to levels 0–2, while F_{cls} bypasses fusion under Pure Dense Strategy.
Training & Hyperparameters. We implement our method using the MMDetection codebase [3]. All models are trained on 6 GPUs. *Optimization.* We use the AdamW [24] optimizer with a generic learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. Both backbones are frozen and we add only **3M** trainable parameters. The learning rate for the backbone is set to 0 (frozen), while the learning rate for the newly introduced fusion modules is set to 1×10^{-4}. *Schedule.* For the OV-COCO benchmark, we train the model for 30 epochs with a total batch size of 6. For the OV-LVIS benchmark, due to the larger dataset size, we extend the training to 35 epochs. *Loss Functions.* We adopt the standard Hungarian matching loss coefficients: $\lambda_{cls} = 2.0$, $\lambda_{L1} = 5.0$, and $\lambda_{GIoU} = 2.0$. To stabilize the training process, we employ EMA with a decay rate of 0.999.

4.3 Benchmark Results

Results on OV-COCO. Tab. 1(a) summarizes the performance comparison on the OV-COCO benchmark. Our proposed ESSA-OVD sets a new state-of-the-art across different backbones. Specifically, with the ViT-L/14 backbone, ESSA-OVD achieves a breakthrough performance of **47.6** AP_{50} on novel categories. This result represents a substantial improvement over existing methods. First, it surpasses the equivalent backbone competitor CLIPSelf [35] (with ViT-L/14) by a remarkable margin of **3.3** AP_{50}. More importantly, ESSA-OVD outperforms the current state-of-the-art method CCKT-Det++ [41]. Even when comparing against CCKT-Det++ variants equipped with ResNet-50 [11] and Swin-B [23] backbones (which achieve 45.3 and 46.0 AP_{50} respectively), our method still leads by significant margins of **2.3** and **1.6** AP_{50} without using MLLM. This establishes a new performance ceiling for the task, demonstrating that our explicit semantic-spatial alignment strategy is superior to simply scaling up the backbone capacity. In addition to the large model, our method with the ViT-B

Table 1: Comparisons on OV-COCO and OV-LVIS benchmarks. Ours outperforms state-of-the-art methods.Image description supervision means that the method learns from additional image-text pairs, while CLIP supervision refers to transferring knowledge from CLIP.

(a) OV-COCO benchmark

Method	Supervision	Backbone	AP_{50}^{Novel}
ViLD [8]	CLIP	RN50	27.6
Detic [45]	Caption	RN50	27.8
OV-DETR [38]	CLIP	RN50	29.4
RTGen [2]	Caption	RN50	33.6
BARON [33]	CLIP	RN50	34.0
CLIM [34]	CLIP	RN50	36.9
SAS-Det [43]	CLIP	RN50	37.4
OV-DQUO [32]	CLIP	RN50	39.2
CCKT-Det++ [41]	CLIP	RN50	45.3
RegionCLIP [44]	Captions	RN50x4	39.3
CORA [36]	CLIP	RN50x4	41.7
OV-DQUO [32]	CLIP	RN50x4	45.6
PromptDet [6]	Caption	ViT-B/16	30.6
CLIPSelf [35]	CLIP	ViT-B/16	<u>37.6</u>
RO-ViT [16]	CLIP	ViT-L/16	33.0
CFM-ViT [15]	CLIP	ViT-L/16	34.1
BIND [42]	CLIP	ViT-L/16	41.5
CLIPSelf [35]	CLIP	ViT-L/14	<u>44.3</u>
CCKT-Det++ [41]	CLIP	SwinB	<u>46.0</u>
ESSA-OVD (Ours)	CLIP	ViT-B/16	**40.0**
ESSA-OVD (Ours)	CLIP	ViT-L/14	**47.6**

(b) OV-LVIS benchmark

Method	Supervision	Backbone	mAP_r
ViLD [8]	CLIP	RN50	16.3
OV-DETR [38]	CLIP	RN50	17.4
BARON [33]	CLIP	RN50	22.6
RegionCLIP [44]	Caption	RN50x4	22.0
$CORA^+$ [36]	Caption	RN50x4	28.1
SAS-Det [43]	CLIP	RN50x4	29.1
CLIM [34]	CLIP	RN50x64	32.3
F-VLM [19]	CLIP	RN50x64	32.8
RTGen [2]	Caption	Swin-B	30.2
BIND [42]	CLIP	ViT-L/16	32.5
Detic [45]	Caption	Swin-B	33.8
CFM-ViT [15]	CLIP	ViT-L/14	33.9
RO-ViT [16]	CLIP	ViT-H/16	34.1
CLIPSelf [35]	CLIP	ViT-L/14	34.9
ProxyDet [12]	Caption	Swin-B	36.7
CoDet [25]	Caption	ViT-L/14	37.0
OV-DQUO [32]	CLIP	ViT-B/16	<u>29.7</u>
OV-DQUO [32]	CLIP	ViT-L/14	<u>39.3</u>
ESSA-OVD (Ours)	CLIP	ViT-B/16	**32.8**
ESSA-OVD (Ours)	CLIP	ViT-L/14	**41.3**

backbone also demonstrates competitive performance, verifying that our strategy is scalable across different model capacities.

Results on OV-LVIS. We further validate ESSA-OVD on the large-vocabulary LVIS benchmark, which is characterized by its long-tailed object distribution. As shown in Tab. 1(b), our method achieves superior performance on the most challenging Rare categories (AP_r). Using the ViT-L and ViT-B backbone, ESSA-OVD attains **41.3** and **32.8** AP_r, which surpasses the equivalent backbone sota OV-DQUO [32] by **2.0** and **3.1** AP_r. Crucially, our top-tier performance with ViT-L significantly outperforms recent state-of-the-art competitors. For instance, it surpasses CLIPSelf [35] (34.9 AP_r) and CoDet [25] (37.0 AP_r) by remarkable margins of **6.4** and **4.3** AP_r, respectively. It is worth noting that our method even outperforms RO-ViT [16] equipped with a significantly larger ViT-H backbone (34.1 AP_r). This success is primarily attributed to our Semantic-Spatial Alignment and Frequency-Aware Fusion and Spatial Adapter modules, which successfully leverage high-frequency spatial priors from DINOv2 to com-

Table 2: Comparisons of the LVIS-trained model cross evaluated on the COCO [22] and Objects365 [30] datasets are presented without any fine-tuning

Methods	**COCO** [22]			**Objects365** [30]		
	AP (%)	AP_{50} (%)	AP_{75} (%)	AP (%)	AP_{50} (%)	AP_{75} (%)
Supervised	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [8]	36.6	55.6	39.8	11.8	18.2	12.6
DetPro [5]	34.9	53.8	37.4	12.1	18.8	12.9
F-VLM [19]	32.5	53.1	34.6	11.9	19.2	12.6
BARON [33]	36.2	55.7	39.1	13.6	21.0	14.5
CCKT-Det++ [41]	38.5	53.2	42.1	15.2	20.9	15.8
ESSA-OVD (Ours)	**40.0**	54.5	**43.3**	**17.1**	**23.9**	**18.1**

pensate for the localized semantic blurring often found in standard CLIP-based features.

Cross-Dataset Evaluation. To further verify the generalization capability of ESSA-OVD in real-world open-world scenarios, we conduct cross-dataset evaluations by directly applying the model trained on OV-LVIS with the ViT-L/14 backbone to the validation sets of **OV-COCO** [22] and **Objects365** [30], without any fine-tuning. As shown in Tab. 2, our method exhibits strong transferability. On OV-COCO, ESSA-OVD achieves a competitive **40.0** AP, demonstrating that it learns robust object representations rather than overfitting to the LVIS distribution. More importantly, on the challenging Objects365 dataset containing 365 diverse categories, our method attains **17.1** AP, surpassing the previous state-of-the-art CCKT-Det++ by **1.9** AP. These results confirm that our explicit alignment strategy effectively learns universal semantic–spatial correspondences, enabling the detector to handle diverse object concepts across different domains.

4.4 Ablation Study

In this subsection, we dismantle our proposed ESSA-OVD framework to analyze the contribution of each individual component on the OV-LVIS benchmark, as summarized in Tab. 3(a).

Effect of Semantic-Spatial Alignment. The baseline model, relying solely on the CLIP image encoder, suffers from the semantic-spatial discrepancy, yielding a limited 38.1 mAP_r. By introducing the auxiliary visual stream to enable explicit Semantic-Spatial Alignment, we observe a substantial performance leap of **1.5** mAP_r. This validates our core premise: the coarse-grained semantic features of CLIP are insufficient for precise localization, and infusing object-centric visual priors is a fundamental prerequisite for accurate OVD.

Effect of Spatial Adapter. Simply introducing auxiliary features is not enough due to the feature space mismatch between the semantic (CLIP) and spatial (DI-

Table 3: Ablation studies on OV-LVIS benchmark.

(a) The ablation study results by using different combinations of Semantic-Spatial-Alignment, Spatial-Adapter and Fre-Aware-Fusion parts.

#	Semantic-Spatial Alignment	Spatial Adapter	Fre-Aware Fusion	mAP_r
1	-	-	-	38.1
2	✓	✗	✗	39.6
3	✓	✓	✗	40.3
4	✓	✗	✓	40.5
5	✓	✓	✓	**41.3**

(b) The ablation study results by using different architectures for both Semantic backbone and Spatial Alignment part backbone.

#	Semantic Backbone	Spatial-Alignment Backbone	mAP_r
1	ViT-L-14	ViT-L-14	**41.3**
2	ViT-L-14	ViT-B-14	40.8
3	ViT-B-16	ViT-L-14	32.8
4	ViT-B-16	ViT-B-14	32.2

(c) Effect of different Adapter performance.

#	Spatial Adapter	CBAM Adapter	LFA Adapter	mAP_r
1	-	-	-	40.5
2	✗	✗	✓	40.6
3	✗	✓	✗	41.0
4	✓	✗	✗	**41.3**

(d) Effect of different Fuse part performance.

#	Gated Fusion	Channel Fusion	Fre-Aware Fusion	mAP_r
1	-	-	-	40.3
2	✓	✗	✗	40.4
3	✗	✓	✗	40.1
4	✗	✗	✓	**41.3**

NOv2) encoders. As shown in Tab. 3(a), incorporating the Spatial Adapter further boosts the performance to 40.3 mAP_r. This improvement indicates that the adapter effectively recalibrates the auxiliary features, suppressing background noise and aligning the spatial cues with the detector's semantic query space. Furthermore, as detailed in Tab. 3(c), our Spatial Adapter outperforms the LFA Adapter, likely because it directly preserves explicit spatial structures without introducing spectral artifacts. Moreover, our design not only achieves superior performance compared to the highly complex CBAM Adapter (41.0 mAP_r), but also requires significantly **fewer parameters**, ensuring much **higher computational efficiency**.

Effect of Frequency-Aware Fusion. Finally, we integrate the Fre-Aware Fusion module. This mechanism, which selectively injects high-frequency spectral components, brings the performance to a peak of **41.3** mAP_r. Comparing Tab. 3(a) line 3 and line 5, the gain demonstrates that while the adapter aligns the features globally, the frequency-aware fusion is crucial for recovering fine-grained boundary details lost during downsampling. Notably, applying fusion alone or without proper adaptation yields suboptimal results compared to the full pipeline, confirming that our staged alignment strategy—first align, then infuse—creates a strong synergistic effect for open-vocabulary detection. As validated in Tab. 3(d), our module also surpasses alternative designs. Specifically, standard Gated Fusion yields only marginal gains (40.4 mAP_r) since it operates in the spatial domain and struggles to decouple structural details from conflicting semantic noise. Moreover, Channel Fusion even degrades performance

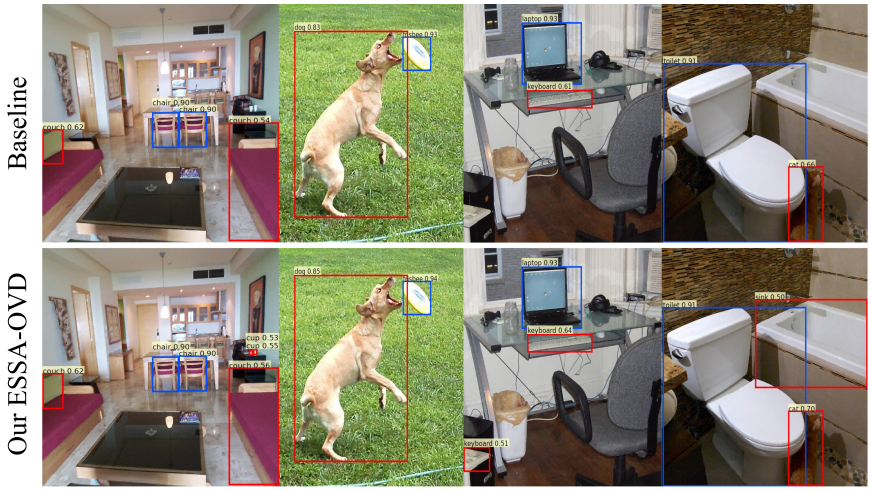

Fig. 3: Visualization of detection results for the OV-COCO benchmark compared with the baseline model. Red boxes represent novel categories and blue boxes indicate base categories.

to 40.1 mAP_r, likely because optimizing attention across massive channel dimensions dilutes the learning focus and causes mutual interference. Our explicitly targeted frequency design successfully avoids both pitfalls.

Analysis of Semantic–Spatial Scaling. In Tab. 3(b), we vary the model scales of the semantic backbone (CLIP) and the spatial source (DINOv2) to disentangle the gains from model capacity and from our Semantic–Spatial Alignment. We observe that the alignment consistently improves performance across both semantic backbone sizes (ViT-L/14 and ViT-B/16), indicating that the proposed framework is not tied to a specific scale. More importantly, reducing the spatial source from ViT-L/14 to a smaller ViT-B/14 incurs only a minor drop (41.3→40.8 with ViT-L/14 backbone, and 32.8→32.2 with ViT-B/16 backbone), while still maintaining strong performance compared to the corresponding baselines. This suggests that the improvements mainly stem from effective semantic–spatial alignment rather than simply increasing the parameter count, and also provides a cost-effective choice for practical deployment.

These comparisons allow us to decouple the contributions of semantic understanding and spatial localization, providing a guideline for selecting the optimal trade-off between accuracy and inference cost.

4.5 Visualization Results

OV-COCO Visualization Results. In Fig. 3, we compare our method with a baseline on OV-COCO. Our approach yields more accurate detections across base and novel categories, especially for unseen concepts where the baseline often misses or mislocalizes objects. The sharper boundaries support the effec-

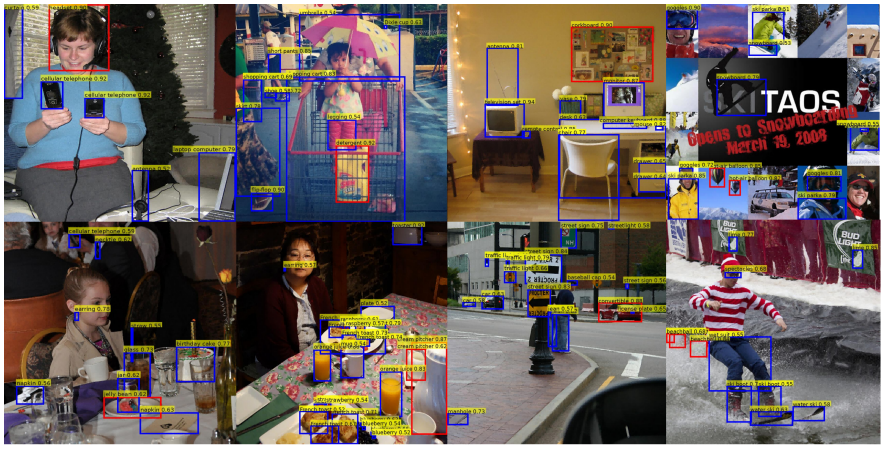

Fig. 4: Visualization of detection results for the OV-LVIS benchmark. Red boxes represent rare categories and blue boxes indicate common and frequent categories.

tiveness of our Explicit Semantic-Spatial Alignment in recovering spatial details.
OV-LVIS Visualization Results. Furthermore, in Figure 4, we visualize the detection results on the more challenging and large scale OV-LVIS [9] benchmark. The visualizations confirm that our framework can effectively recognize a wide spectrum of rare categories alongside common and frequent categories in complex real-world scenes. This further highlights the robust base-to-novel generalization capability of our method without sacrificing the detection performance of known classes.

5 Conclusion

In this paper, we propose ESSA-OVD to address the inherent semantic–spatial discrepancy in open-vocabulary object detection. Its dual-stream architecture explicitly aligns CLIP's semantic generalization with DINOv2's high-frequency structural priors. Unlike methods relying on extra supervision, ESSA-OVD intrinsically recovers missing spatial spectra while preventing semantic drift through decoupled feature routing. Extensive experiments demonstrate state-of-the-art performance across multiple benchmarks. Future work will explore more efficient and lightweight alignment strategies for broader open-world perception.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62573163 and 62503139, in part by the GuangDong Basic and Applied Basic Research Foundation under Grants 2024A1515012028 and 2026A1515011822, and in part by the Shenzhen Science and Technology Program under Grants JCYJ20250604145514019 and GXWD20231129125006001.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
2. Chen, F., Zhang, H., Yang, Z., Chen, H., Hu, K., Savvides, M.: Rtgen: Generating region-text pairs for open-vocabulary object detection. arXiv preprint arXiv:2405.19854 (2024)
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
4. Choi, J., Lee, S., Lee, M.S., Lee, S., Shim, H.: Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In: CVPR. pp. 9782–9793 (2025)
5. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR. pp. 14064–14073 (2022)
6. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: ECCV. pp. 701–717. Springer (2022)
7. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. IJCV **132**(2), 581–595 (2024)
8. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022)
9. Gupta, A., Dollár, P., Girshick, R.B.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019)
10. Han, X., Wei, L., Yu, X., Dou, Z., He, X., Wang, K., Sun, Y., Han, Z., Tian, Q.: Boosting segment anything model towards open-vocabulary learning. In: AAAI. vol. 39, pp. 3356–3365 (2025)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. Jeong, J., Park, G., Yoo, J., Jung, H., Kim, H.: ProxyDet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In: AAAI. vol. 38, pp. 2462–2470 (2024)
13. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PmLR (2021)
14. Kim, C., Ju, D., Han, W., Yang, M., Hwang, S.J.: Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In: CVPR. pp. 15033–15042 (2025)
15. Kim, D., Angelova, A., Kuo, W.: Contrastive feature masking open-vocabulary vision transformer. In: ICCV. pp. 15602–15612 (2023)
16. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: CVPR. pp. 11144–11154 (2023)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
18. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: ICLR (2022)

19. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-VLM: Open-vocabulary object detection upon frozen vision and language models. In: ICLR (2023)
20. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted CLIP. In: CVPR. pp. 7061–7070 (2023)
21. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: ICLR (2023)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In: NeurIPS. vol. 36, pp. 71078–71094 (2023)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
28. Redmon, J., Divvala, S., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
29. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS. vol. 28, pp. 91–99 (2015)
30. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV. pp. 8430–8439 (2019)
31. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
32. Wang, J., Chen, B., Kang, B., Li, Y., Xian, W., Chen, Y., Xu, Y.: OV-DQUO: Open-vocabulary DETR with denoising text query training and open-world unknown objects supervision. In: AAAI. vol. 39, pp. 7762–7770 (2025)
33. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: CVPR. pp. 15254–15264 (2023)
34. Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., Loy, C.C.: CLIM: contrastive language-image mosaic for region representation. In: AAAI. vol. 38, pp. 6117–6125 (2024)
35. Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., Loy, C.C.: Clipself: Vision transformer distills itself for open-vocabulary object detection. In: ICLR (2024)
36. Wu, X., Zhu, F., Zhao, R., Li, H.: CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In: CVPR. pp. 7031–7040 (2023)
37. Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K., Ren, F.: Learning in the frequency domain. In: CVPR. pp. 1737–1746 (2020)
38. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary DETR with conditional matching. In: ECCV. pp. 106–122. Springer (2022)

39. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR. pp. 14393–14402 (2021)
40. Zhai, H., Cheng, J., Wang, M.: Rethink the IoU-based loss functions for bounding box regression. In: ITAIC. vol. 9, pp. 1522–1528. IEEE (2020)
41. Zhang, C., Zhu, C., Dong, P., Chen, L., Zhang, D.: Cyclic contrastive knowledge transfer for open-vocabulary object detection. In: ICLR (2025)
42. Zhang, H., Zhao, Q., Zheng, L., Zeng, H., Ge, Z., Li, T., Xu, S.: Exploring region-word alignment in built-in detector for open-vocabulary object detection. In: CVPR. pp. 16975–16984 (2024)
43. Zhao, S., Schulter, S., Zhao, L., Zhang, Z., Kumar, B.G.V., Suh, Y., Chandraker, M., Metaxas, D.N.: Taming self-training for open-vocabulary object detection. In: CVPR. pp. 13938–13947 (2024)
44. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16772–16782 (2022)
45. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV. pp. 350–368. Springer (2022)
46. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)