This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhancing Novel Object Detection via Cooperative Foundational Models

Rohit Bharadwaj[1] Muzammal Naseer[4] Salman Khan[1,3] Fahad Shahbaz Khan[1,2]
[1]MBZUAI [2]Linköping University [3]Australian National University [4]Khalifa University
rohit.bharadwaj@mbzuai.ac.ae, muhammadmuzammal.naseer@ku.ac.ae,
salman.khan@mbzuai.ac.ae, fahad.khan@mbzuai.ac.ae

Abstract

In this work, we address the challenging and emergent problem of novel object detection (NOD), focusing on the accurate detection of both known and novel object categories during inference. Traditional object detection algorithms are inherently closed-set, limiting their capability to handle NOD. We present a novel approach to transform existing closed-set detectors into open-set detectors. This transformation is achieved by leveraging the complementary strengths of pre-trained foundational models, specifically CLIP and SAM, through our cooperative mechanism. Furthermore, by integrating this mechanism with state-of-the-art open-set detectors such as GDINO, we establish new benchmarks in object detection performance. Our method achieves 17.42 mAP in novel object detection and 42.08 mAP for known objects on the challenging LVIS dataset. Adapting our approach to the COCO OVD split, we obtain an impressive result of 49.6 Novel AP50, which outperforms existing SOTA methods with similar backbone. Our code is available at:
https://rohit901.github.io/coop-foundation-models/

1. Introduction

Object detection serves as a cornerstone in computer vision, with broad applications from autonomous driving and robotic vision to video surveillance and pedestrian detection [3, 29, 41]. Current state-of-the-art approaches such as Mask-RCNN [11], and DETR [2] operate under a closed-set paradigm, where detection is limited to predefined classes seen during training. This limitation does not align with the dynamic and evolving nature of real-world environments where object classes follow a long-tail distribution, with numerous rare and a few common classes. Collecting resource-intensive datasets to represent these rare classes is an impractical task [10]. Thus, the development of open-set detectors that can generalize beyond their training data is crucial for their practical deployment.

The challenge of detecting novel objects is exacerbated by the absence of labeled annotations for such classes. Prior efforts have largely addressed *Generalized Novel Class Discovery (GNCD)*, utilizing semi-supervised and contrastive learning techniques that assume access to pre-cropped objects and balanced datasets. However, these methods are predominantly focused on classification rather than localization, leaving a gap in novel object detection capabilities. RNCDL[7] is a pioneering approach addressing novel object detection (NOD; requires both localization and recognition of known and novel classes) through a training pipeline encompassing both supervised and self-supervised learning. Nonetheless, RNCDL lacks the ability to assign semantic labels to novel categories without additional validation sets, and its performance falls short for practical applications.

(a) RNCDL output (b) Our Mask-RCNN method

Figure 1. Comparison of top-10 predictions: (a) RNCDL [7] can result in imprecise localization and misclassification (e.g., basket, apple), versus (b) our open-set Mask-RCNN, demonstrating accurate detection and categorization of unique objects in the scene.

Emerging approaches in *open-vocabulary object detection (OVD; similar to NOD, but main focus is on novel classes, and requires large-scale training)* have demonstrated promise through the integration of Vision-Language Models (VLMs) like CLIP [24], which leverage language embeddings of category names to generalize detection capabilities. Techniques such as GLIP [15] and Grounding DINO (GDINO) [19] further intertwine language and vision modalities at various architectural stages, achieving detection through text inputs. These advancements, however, are contingent upon extensive training and computational resources, implicating significant environmental and financial costs [22, 27].

In this work, we present a cooperative mechanism that harnesses the complementary strengths of foundational models such as CLIP and SAM to transition existing closed-set detectors into open-set detectors for novel object detection (NOD). These foundational models, trained on diverse datasets, are adept at generalizing across tasks and distributions unseen during training. Our approach uses CLIP's zero-shot classification in tandem with background boxes identified by Mask-RCNN to determine novel class labels and their confidence scores. These boxes then guide SAM in refining the predictions and eliminating spurious background detections. Ours is the first approach to demonstrate the potential of existing foundational VLMs for NOD.

We further demonstrate that our cooperative mechanism, when paired with open-set detectors like GDINO, elevates performance metrics across both known and novel object categories. Through extensive evaluations on the challenging LVIS [10] dataset and the COCO [18] open-vocabulary benchmark, our method sets new state-of-the-art, particularly in Novel AP50 metric. Ablation studies highlight the significant contributions of individual and novel components, such as the Synonym Averaged Embedding Generator (SAEG) and the Score Refinement Module (SRM). Our contributions are as follows:

1. A cooperative mechanism leveraging the strengths of pre-trained foundational models like CLIP and SAM, transforming closed-set detectors into proficient open-set detectors for novel object detection.

2. A modular design enabling integration with state-of-the-art open-set detectors to enhance Known and Novel class AP in challenging settings.

3. We demonstrate the effectiveness of our approach compared to existing state-of-the-art in open-set, unknown, and open-vocabulary object detection by various experiments conducted on challenging datasets like COCO and LVIS. Further, our ablation studies demonstrate the strength of proposed approach and its constituent elements.

2. Related Work

Recent research in object detection has made significant strides, yet the challenge of detecting objects outside the predefined classes remains unsolved. Existing methods often assume a semi-supervised setting and focus on classification rather than the complex task of simultaneous classification and localization, which our work tackles. GCD [28] exploits the capabilities of Vision Transformers and contrastive learning to identify novel classes, avoiding the overfitting issues of parametric classifiers. OpenLDN [26] adopts a similar goal, leveraging image similarities and bi-level optimization to cluster novel class instances, while PromptCAL [39] introduces contrastive affinity learning with visual prompts to enhance class discovery. In the realm of unknown object detection, methods such as VOS [5] synthesize virtual outliers to differentiate between known and unknown objects. UnSniffer [16], the current leading method, uses a generalized object confidence score and an energy suppression loss to distinguish non-object background samples. For Open-Vocabulary Detection (OVD), approaches like ViLD [9], OV-DETR[35], and BARON [32] employ knowledge distillation from Vision-Language Models to align region embeddings with VLM features. The state-of-the-art CORA[33] further refines this by combining region prompting with anchor prematching in a DETR-based framework. RNCDL [7] also addresses novel object detection, but its two-stage training pipeline, reliance on a validation set for semantic label assignment, and poor performance limit its practicality. Unlike RNCDL, our work is able to achieve superior performance in a training-free manner. Lastly, methods like GLIP [15] and GDINO [19] harness natural language to expand detection capabilities. However, the extensive training required for these methods incurs significant financial and environmental costs. Our approach, on the other hand, leverages complementary strengths of pre-trained foundational models to achieve superior performance without the need for extensive training, showcasing a more practical and accessible solution for novel object detection. Further, the modular nature of our cooperative mechanism allows it to be combined with any existing state-of-the-art open-set detectors like GDINO to further improve the overall performance across both known and novel classes.

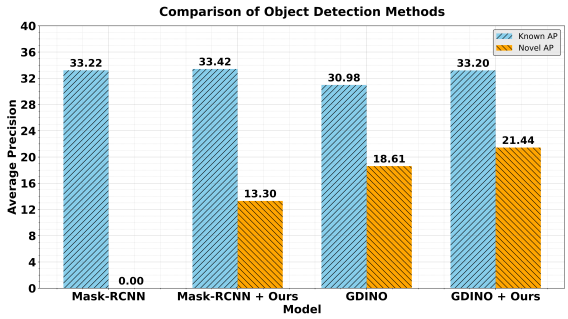


Figure 2. Comparative analysis of object detection methods on `lvis_v1_val_subset` dataset. The closed-set Mask-RCNN does not detect novel classes, however, the performance consistently improves when combined with our cooperative mechanism integrating different foundational models.

3. Novel Object Detection

Novel Object Detection (NOD) deals with object detection under heterogeneous object distributions during train-

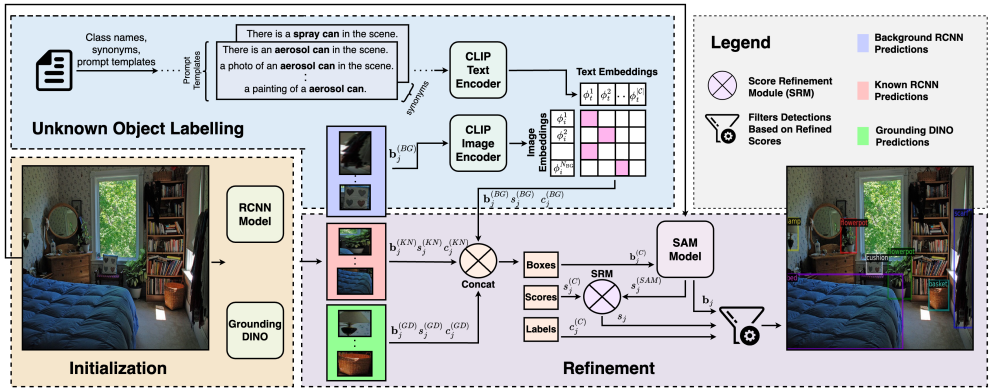


Figure 3. Our proposed cooperative mechanism integrates pre-trained foundational models such as CLIP, SAM, and GDINO with a Mask-RCNN model in order to identify and semantically label both known and novel objects. These foundational model interacts using different components including Initialization, Unknown Object Labelling, and Refinement to refine and categorize objects.

ing and inference. Specifically, we assume that the distribution of classes observed during training might differ from the distribution observed during inference. As a result, we may encounter both known and previously unknown object classes. Our objective is to identify objects from known and novel categories during inference while assigning semantic labels to each object.

Formally, we define $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}_{\text{train}}$ as the training dataset. Each $\mathbf{x}_i \in \mathbb{R}^{3\times h\times w}$ represents an input image and $y_i = \{(\mathbf{b}_j, c_j)\}_{j=1}^{L}$ comprises bounding box label $\mathbf{b}_j \in \mathbb{R}^4$ and class label $c_j \in \mathcal{C}^{\text{known}}$. We represent the test dataset as $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \in \mathcal{X} \times \mathcal{Y}_{\text{test}}$, where $\mathcal{Y}_{\text{test}} \supset \mathcal{Y}_{\text{train}}$. Importantly, class labels during testing, c_j, can belong to an extended set $\mathcal{C} \supset \mathcal{C}^{\text{known}}$.

Novel classes are by definition unknown; therefore, it is challenging to detect them during inference. The conventional object detection models such as Mask-RCNN [11] and DETR [2] are examples of closed-set detectors unable to generalize beyond the classes of their training data. These models are unsuitable for detecting novel objects. We therefore require open-set object detectors to solve the problem of NOD. Recently, Grounding DINO [19] (GDINO) achieved open-set generalization by introducing natural language to a closed-set detector DINO [38]. The model processes natural language input in the form of category names or classes to detect arbitrary objects within an image. The key to the open-set success of GDINO is language and vision fusion along with large-scale training.

Approach Overview. In this work, we show how to convert an existing closed-set detector, i.e, pre-trained Mask-RCNN, to an open-set detector by utilizing the complementary strengths of pre-trained foundational models such as CLIP [24], and SAM [14] via our *cooperative mechanism*. Our proposed mechanism leverages CLIP's understanding of unseen classes with Mask-RCNN's ability to localize background objects for finding novel object classes (Sec. 3.2). The bounding boxes are then refined by using SAM's instance mask-to-box capabilities (Algorithm 1). We observe that the novel class AP of Mask-RCNN is zero due to its closed-set nature (Fig. 2). Our cooperative mechanism, however, can be applied to the same closed-set Mask-RCNN model and result in notable gains in novel AP, as well as better performance in known AP compared to GDINO and the baseline Mask-RCNN. Additionally, when our proposed cooperative mechanism is combined with open-set detectors like GDINO, we observe overall performance gains (Fig. 2).

Our approach uses off-the-shelf pre-trained models in three stages; a) Initialization, b) Unknown Object Labelling, and c) Refinement. The initialization stage consists of getting bounding boxes of known and unknown object categories from complementary off-the-shelf detectors (Mask-RCNN [11], Grounding DINO [19]). The second stage involves processing unlabelled background bounding boxes by a pre-trained language-image model CLIP [24]. Finally, the detected boxes are refined using SAM [14]. We first provide an overview of CLIP, SAM, and GDINO in the next Sec. 3.1.

3.1. Preliminaries

Contrastive Language-Image Pre-training (CLIP). The CLIP model incorporates a dual-encoder architecture tailored for multi-modal learning, where the text and image encoders are represented by $\mathcal{F}_t^{(\text{CLIP})}$ and $\mathcal{F}_i^{(\text{CLIP})}$, respectively. Training involves a batch of N image-text pairs, where the goal is to align the image embeddings $\boldsymbol{\phi}_i \in \mathbb{R}^d$ with their corresponding text embeddings $\boldsymbol{\phi}_t \in \mathbb{R}^d$.

For zero-shot classification, given an image $\mathbf{x}$ and a set of $|\mathcal{C}|$ unique classes, we generate $|\mathcal{C}|$ textual prompts using a template $T(\cdot)$ — for instance, "a photo of a [CLASS]." By

replacing the [CLASS] token with each class name c from our targeted dataset, we form the textual descriptors. The embeddings for these descriptors, $\mathbf{\Phi}_t = \mathcal{F}_t^{(\text{CLIP})}(T(\mathcal{C}))$, and for the image $\boldsymbol{\phi}_i = \mathcal{F}_i^{(\text{CLIP})}(\mathbf{x})$, are used to compute logits as: $\boldsymbol{\phi}_i^T \mathbf{\Phi}_t \in \mathbb{R}^{1\times|\mathcal{C}|}$. The predicted class for the image is the one with the highest corresponding score.

Segment Anything Model (SAM). SAM is a recent foundational model for generic class-agnostic image segmentation. The model is comprised of three primary modules: an image encoder $\mathcal{F}_i^{(\text{SAM})}$, a prompt encoder $\mathcal{F}_p^{(\text{SAM})}$, and a mask decoder $\mathcal{G}_m^{(\text{SAM})}$. Formally, for an input image $\mathbf{x}$, and a set of prompts $\mathcal{P} = \{\mathbf{p}_1, ..., \mathbf{p}_M\}$, SAM generates refined segmentation masks $\{\mathbf{m}_1, ..., \mathbf{m}_M\}$ and a corresponding set of confidence scores $\mathbf{s}^{(\text{SAM})} = \{s_1^{(\text{SAM})}, ..., s_M^{(\text{SAM})}\}$. The prompts $\mathcal{P}$ can include any combination of points, bounding boxes, or rough masks. These are encoded through $\mathcal{F}_p^{(\text{SAM})}$ to produce prompt embeddings $\boldsymbol{\phi}_p$. Concurrently, $\mathcal{F}_i^{(\text{SAM})}$ processes $\mathbf{x}$, yielding image embeddings $\boldsymbol{\phi}_i$. These two embeddings are then fused and passed to $\mathcal{G}_m^{(\text{SAM})}$, for refined segmentation masks $\mathbf{m}$ and their scores $\mathbf{s}^{(\text{SAM})}$.

Grounding DINO (GDINO). GDINO utilizes the DETR [2] architecture for a transformer-based framework for object detection. It adopts the Swin Transformer [20] as the visual backbone and BERT [4] for the textual encoding. In this setup, given a pair consisting of an image $\mathbf{x}$ and a concatenated textual input of class names from the set of all classes $\mathcal{C}$, the model yields a fixed number (`num_query` = 900) of predicted bounding boxes $\{\mathbf{b}_i^{(\text{GD})}\}_{i=1}^{\texttt{num\_query}}$. Alongside, it generates a set of confidence scores $\{s_i^{(\text{GD})}\}_{i=1}^{\texttt{num\_query}}$ and associates each detected region with a class label $\{c_i^{(\text{GD})} \in \mathcal{C}\}_{i=1}^{\texttt{num\_query}}$.

3.2. Cooperative Foundational Models for NOD

Initialization In our proposed pipeline, the first stage involves initializing bounding boxes using off-the-shelf detectors, such as Mask-RCNN and GDINO, to obtain the unrefined bounding boxes from the input image. The outputs of the detectors based on a conventional two-stage RPN design are combined with a complementary DETR architecture to compensate for the weaknesses of each component. For example, Mask-RCNN does not integrate linguistic cues, and DETR has only a limited set of predicted bounding boxes (i.e. `num_query`).

Specifically, the Mask-RCNN is a unimodal, non-transformer based two-stage model, hence it does not provide the flexibility to detect objects from an open-vocabulary. The RCNN family of object detection models are known as two-stage detectors because, in the first stage, they output all the object proposals in the given image, which may include known as well as unidentifiable background objects, and in the second stage, the model generates the final box predictions based on the initial set of proposals. In the absence of class labels for novel objects within a dataset, these categories are generally classified as *background* class. To solve the NOD problem, the Mask-RCNN model needs to output bounding boxes for background classes as well as known classes. Similarly, due to the `num_query` bounding box limitation, the DETR based GDINO model does not perform well on rare classes e.g., within the LVIS dataset [19]. As shown in Fig. 3, GDINO misses out on detecting some of the other objects like *cushion*, and *scarf* which are present in the input image.

Formally, for a given input image, $\mathbf{x}$, we obtain the outputs from GDINO - specifically, the bounding boxes $\mathbf{b}_j^{(\text{GD})}$, class confidence scores $s_j^{(\text{GD})}$, and the associated class label $c_j^{(\text{GD})} \in \mathcal{C}$, where $j = 1, \ldots, N_{\text{GD}}$. For the same input image $\mathbf{x}$, we also get the Mask-RCNN outputs which include the known bounding boxes; $\mathbf{b}_j^{(\text{KN})}$, class confidence scores $s_j^{(\text{KN})}$, and class labels $c_j^{(\text{KN})} \in \mathcal{C}^{\text{known}}$, where $j = 1, \ldots, N_{\text{KN}}$, and the background bounding boxes; bounding boxes $\mathbf{b}_j^{(\text{BG})}$, where $j = 1, \ldots, N_{\text{BG}}$. Since the closed-set Mask-RCNN is not able to output the class labels and confidence scores for the background boxes, we describe our approach utilizing CLIP to obtain this missing data, and to convert the closed-set Mask-RCNN model to an open-set detector in the next subsection.

Unknown Object Labelling To convert the existing closed-set Mask-RCNN to an open-set detector, we utilize the zero-shot capabilities of a vision language model such as CLIP to obtain the class labels and confidence scores for the background boxes $\mathbf{b}_j^{(\text{BG})}$. First, the regions of interest (ROIs) are cropped from the raw images, $\mathbf{x}_i$, using the background boxes, $\mathbf{b}_j^{(BG)}$. We obtain N_{BG} number of ROIs that serve as the image inputs for zero-shot classification with CLIP. As shown in Fig. 3, we obtain the visual embedding for these ROIs denoted as $\{\boldsymbol{\phi}_i^k\}_{k=1}^{N_{BG}}$.

Synonym Averaged Embedding Generator (SAEG). To obtain the text embeddings corresponding to each class $c \in \mathcal{C}$, we define $\mathcal{T}$ as set of all prompt templates, $\mathcal{S}_i$ as set of all synonyms for the i-th class C_i in the dataset and $\mathcal{C}$ as set of all classes, $\{C_1, C_2, \ldots, C_{|\mathcal{C}|}\}$. Then for each synonym s in $\mathcal{S}_i$, we proceed as follows. First, the set of prompted texts for the synonym s is generated as $\mathcal{P}_s = \{T(s) : T \in \mathcal{T}\}$. Afterwards, these prompted texts are tokenized and encoded using CLIP to obtain their embeddings, i.e, $\boldsymbol{\phi}_t^{\text{syn}} = \mathcal{F}_t^{(\text{CLIP})}(\mathcal{P}_s) \in \mathbb{R}^{d\times n}$, where $n = |\mathcal{P}_s|$. The resulting embeddings are normalized and averaged to obtain the representative text feature $\mathbf{f}_s \in \mathbb{R}^d$ for the synonym s.

$$\mathbf{f}_s = \frac{1}{|\mathcal{P}_s|} \sum_{\mathbf{e}\in\boldsymbol{\phi}_t^{\text{syn}}} \frac{\mathbf{e}}{\|\mathbf{e}\|}. \tag{1}$$

The text features for all synonyms of a class are averaged to generate the text feature $\boldsymbol{\phi}_t^i \in \mathbb{R}^d$ for that class C_i.

$$\boldsymbol{\phi}_t^i = \frac{1}{|\mathcal{S}_i|} \sum_{\mathbf{f}_s \in \mathcal{S}_i} \frac{\mathbf{f}_s}{\|\mathbf{f}_s\|}. \tag{2}$$

The final text feature matrix $\boldsymbol{\Phi}_t$ for all possible classes can then be represented as:

$$\boldsymbol{\Phi}_t = [\boldsymbol{\phi}_t^1, \boldsymbol{\phi}_t^2, \ldots, \boldsymbol{\phi}_t^{|\mathcal{C}|}] \in \mathbb{R}^{d \times |\mathcal{C}|}. \tag{3}$$

Cosine similarity between these image and enriched text embeddings is computed to finalize class predictions and their confidence scores for the background boxes.

Refinement We consolidate bounding boxes, confidence scores, and class labels obtained from Mask-RCNN and Grounding DINO models. Specifically, we concatenate the outputs as depicted in Fig. 3 to formulate a unified set of bounding boxes, $\mathbf{b}_j^{(\mathrm{C})}$, confidence scores, $s_j^{(\mathrm{C})}$, and class labels, $c_j^{(\mathrm{C})}$, where $j = 1, \ldots, N_{\mathrm{C}}$. Finally, the total number of combined bounding boxes are $N_{\mathrm{C}} = N_{\mathrm{KN}} + N_{\mathrm{BG}} + N_{\mathrm{GD}}$.

This combined set of bounding boxes, $\mathbf{b}_j^{(\mathrm{C})}$, is inherently noisy. To solve this problem and to achieve robust generalization beyond the training data distribution, we utilize SAM [14], which allows for efficient zero-shot generalization. The raw image $\mathbf{x}_i$, and the set of prompts $\mathcal{P} = \mathbf{b}_j^{(C)}$, serve as inputs to SAM. The output is a set of refined segmentation masks for each of the prompted bounding boxes. From these improved masks, we extract new bounding boxes, constituting our final set of refined bounding box predictions, $\mathbf{b}_j$. Subsequently, we introduce the Score Refinement Module (SRM). This novel component integrates confidence scores obtained from SAM with the combined confidence scores, $s_j^{(\mathrm{C})}$, to filter and re-weight the high-quality object predictions, which significantly improves the overall object detection performance as explained next.

Score Refinement Module (SRM). Our Score Refinement Module (SRM) integrates two sets of scores: mask-quality scores, $s_j^{(\mathrm{SAM})}$, and combined confidence scores $s_j^{(\mathrm{C})}$. Both sets of scores are individually subjected to Min-Max standardization [23]. The final refined confidence scores, s_j, are obtained by element-wise multiplication of the standardized scores. Algorithm 1 describes our SRM for score refinement. Finally, using the refined scores, s_j, refined bounding boxes, $\mathbf{b}_j$, and the combined class labels, $c_j^{(\mathrm{C})}$, we filter high-quality object predictions.

Algorithm 1 Score Refinement Process

Input: Combined scores $s_j^{(\mathrm{C})}$, SAM scores $s_j^{(\mathrm{SAM})}$
Output: Refined scores s_j
Initialize MinMaxScaler (from sk-learn library) for combined scores and SAM scores as scaler, scaler_sam
$s_j^{(\mathrm{C})} \leftarrow$ scaler.fit_transform($s_j^{(\mathrm{C})}$)
$s_j^{(\mathrm{SAM})} \leftarrow$ scaler_sam.fit_transform($s_j^{(\mathrm{SAM})}$)
$s_j = s_j^{(\mathrm{C})} \times s_j^{(\mathrm{SAM})}$

4. Experiments and Results

In Sec. 4.1, we begin by outlining the implementation settings for our study. Following this, in Sec. 4.2, we present our main findings. We compare our work with the current leading models, GDINO [19], and the RNCDL [7]. Since our known and novel class splits on LVIS (80 known; 1123 novel) is more challenging than the conventional LVIS-OVD splits (866 known; 337 novel), we adapt our approach on the COCO OVD split for direct comparison against existing OVD methods in Sec. 4.3. We also report comparisons on the localization capabilities of our method against the latest in *unknown object detection* [16] and *open-set object detection* [19] methods in Sec. 4.4. Our method shows a significant improvement over previous methods. In Sec. 4.5, we report extensive ablation studies. The ablations demonstrate the valuable contributions of each component within our overall proposed approach. We conclude by showing the qualitative visualization of our method in Fig. 6, showing clear improvements with our proposed cooperative mechanism.

4.1. Implementation Settings

Datasets: In our experiments, we primarily utilize two datasets: COCO 2017 [18], and LVIS v1 [10]. The LVIS dataset, while based on the same images from COCO, offers more detailed annotations. It features bounding box and instance segmentation mask annotations for 1,203 classes, encompassing all classes from COCO. The LVIS dataset is divided into 100K training images (`lvis_train`) and 20K validation images (`lvis_val`). Our principal evaluation, as detailed in Sec. 4.2, focuses on the LVIS validation split, consistent with prior work [7]. Additionally, we organize the LVIS validation images in descending order by the count of ground-truth box annotations. From this, we select a subset of 745 images (`lvis_val_subset`) approximately covering all LVIS classes. This subset is used in our ablation studies (Sec. 4.5) and localization-only experiments (Sec. 4.4). In general, our investigations indicate that results on this subset are representative of those on the full validation set. **Known/Novel Split:** Consistent with RNCDL [7], we classify the 80 COCO classes as 'known' and the remaining 1,123 LVIS classes as 'novel'. To align with the COCO OVD split, we categorize 48 COCO classes as known and the rest of the 17 as novel.

Mask-RCNN: We explore two versions of the pre-trained Mask-RCNN model. The first, referred to as "*Mask-RCNN-V1*", adheres to the fully-supervised training methodology

Method	Mask-RCNN	GDINO	VLM	Novel AP	Known AP	All AP
K-Means [21]	-	-	-	0.20	17.77	1.55
Weng et al. [30]	-	-	-	0.27	17.85	1.62
ORCA [1]	-	-	-	0.49	20.57	2.03
UNO [6]	-	-	-	0.61	21.09	2.18
RNCDL [7]	V1	-	-	5.42	25.00	6.92
GDINO [19]	-	✓	-	13.47	37.13	15.30
Ours	V2	✗	CLIP	13.24	**42.61**	15.52
Ours	V1	✓	CLIP	15.37	36.15	16.98
Ours	V1	✓	SigLIP	16.12	37.09	17.74
Ours	V2	✓	SigLIP	**17.42**	42.08	**19.33**

Table 1. Comparison of object detection performance using mAP on the `lvis_val` dataset. The VLM column specifies the Vision-Language Model (VLM) utilized, indicating whether it is CLIP or SigLIP. Overall, the SigLIP version with GDINO and Mask-RCNN-V2 provides the best novel vs. known AP tradeoff.

of RNCDL [7], trained on the $\text{COCO}_{\text{half}}$ dataset to ensure fair comparison. The second version, "*Mask-RCNN-V2*", is trained on the complete COCO dataset, utilizing a ResNet101 [12] based FPN [17] backbone with large scale jittering (LSJ) [8] augmentation. We consider top 300 predictions from Mask-RCNN (i.e. $N_{\text{KN}} + N_{\text{BG}} = 300$). **GDINO:** For our experiments, the Swin-T variant of GDINO is employed, being the only open-sourced variant compatible with our open-vocabulary novel-class set constraint. The `num_query` parameter in GDINO is set to 900, but only the top-300 high-scoring predictions ($N_{\text{GD}} = 300$) are utilized (Fig. 4). **CLIP:** Two variants of CLIP-based models are utilized in our approach. The first is the standard CLIP model with a "`ViT-L/14`" backbone [24], and the second is the recently released SigLIP model [13, 31, 37]. We use the "`ViT-SO400M-14-SigLIP`" backbone from SigLIP, alongside 64 prompt templates from ViLD [9], and synonyms from the LVIS dataset [10] for our methodology. **Evaluation Metrics:** For the the principal results in Sec. 4.2, and the ablation studies in Sec. 4.5, we align with previous studies and report mean average precision (mAP@[0.5:0.95]) [18] across "known", "novel", and "all" classes. Further, following the LVIS protocol [10], we evaluate our method by finally taking top 300 high scoring predictions. For COCO OVD results in Sec. 4.3, we evaluate by taking top 100 high scoring predictions. In Sec. 4.4, recall values at an IOU threshold of 0.5 are evaluated for the localization experiments. For the open-vocabulary comparison in Sec. 4.3, box AP_{50} metrics are reported, following the precedent set by earlier works.

4.2. Comparison with NOD Techniques

The results in Tab. 1 reveal that baseline methods like K-Means and other GNCD approaches demonstrate limited effectiveness in NOD (Novel Object Detection), with Novel AP below 1 mAP. Even RNCDL [7], which involves a sophisticated multi-stage training process and requires extensive hyperparameter tuning, falls short in terms of both Known and Novel AP compared to the GDINO benchmark. In stark contrast, our method, which does not necessitate any training and is conceptually straightforward, significantly outperforms RNCDL. Specifically, with our method using only CLIP, we achieve an improvement of 7.82 mAP in Novel classes and an impressive 17.61 mAP in Known classes over RNCDL. Moreover, by integrating our cooperative mechanism with the GDINO and SigLIP model, we observe additional gains of 3.95 mAP in Novel classes and 4.95 mAP in Known classes compared to the baseline GDINO. The use of advanced VLMs like SigLIP and enhanced pre-trained weights of Mask-RCNN further bolsters the overall object detection performance. This showcases the scalability and modular nature of our approach, reinforcing its potential for diverse NOD applications.

Method	Pre-train	Training	Backbone	**Novel**	Base	All
OVR-CNN [36]	COCO Captions, BERT (BooksCorpus, English Wikipedia)	COCO	ResNet50	22.8	46.0	39.9
ViLD [9]	-	COCO, CLIP	ResNet50	27.6	59.5	51.3
Detic [40]	ImageNet-21K	COCO, ImageNet-21K, Conceptual Captions, CLIP	ResNet50	27.8	47.1	45.0
OV-DETR [35]	-	COCO, CLIP	ResNet50-C4	29.4	61.0	52.7
BARON [32]	SOCO dataset	COCO, CLIP	ResNet50	34	60.4	53.5
CORA [33]	-	COCO, CLIP	ResNet50	35.1	35.5	35.4
Rasheed et al. [25]	MAVL (Flickr30k, COCO, Visual Genome)	COCO, COCO Captions, CLIP	ResNet50	36.6	54.0	49.4
BARON [32]	SOCO dataset, MAVL	COCO, COCO Captions, CLIP	ResNet50-C4	42.7	54.9	51.7
CORA+ [33]	-	COCO, COCO Captions, CLIP	ResNet50x4	43.1	60.9	56.2
DetCLIPv3 [34]	FILIP, Qformer, BERT, CLIP	O365, GoldG, V3Det, GranuCap50M, GranuCap600K, CLIP, InstructBLIP, GPT-4, LLaVA	Swin-T	54.7	42.8	46.9
Ours*	**GDINO (O365,GoldG,Cap4M), SAM (SA-1B), CLIP**	COCO	**ResNet50**	**49.6**	42.4	44.3
Ours*	**GDINO (O365,GoldG,Cap4M), SAM (SA-1B), CLIP**	COCO	**ResNet101**	**50.3**	49.8	49.9

Table 2. Comparison with OVD methods on COCO-OVD data split. "COCO" in the training column refers to the COCO dataset with 48 base class annotations. Our method requires only the closed-set detector to be trained or can be training-free with pre-trained detectors, unlike most existing OVD methods that require explicit training with CLIP or other VLMs. Our method is thus zero-shot, while OVD methods are not (evident from CLIP being present in training column of most methods).

4.3. Comparisons with OVD Methods

Tab. 2 compares our proposed zero-shot approach against various SOTA OVD methods on the COCO-OVD dataset split. Unlike most OVD approaches that require extensive training with VLMs like CLIP, our method leverages only a pre-trained closed-set detector and uses VLMs like CLIP in zero-shot way. Specifically, our method achieves a notable 49.6% AP50 on novel categories, surpassing the majority of the existing methods, including those that rely heavily on VLM training such as CORA+ (43.1%) and BARON MAVL (42.7%). While DetCLIPv3 demonstrates

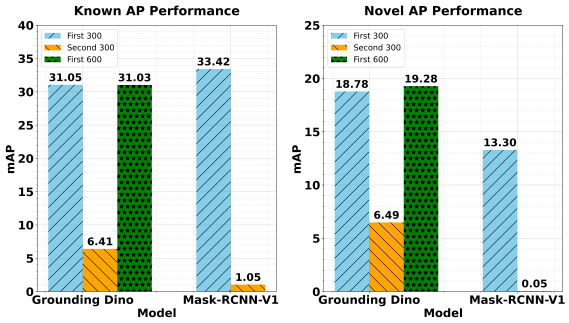


Figure 4. Object detection "mAP" performance on the `lvis_val_subset` dataset. "Grounding Dino" and "Mask-RCNN-V1" models were evaluated after considering the first 300, second 300, and the first 600 boxes.

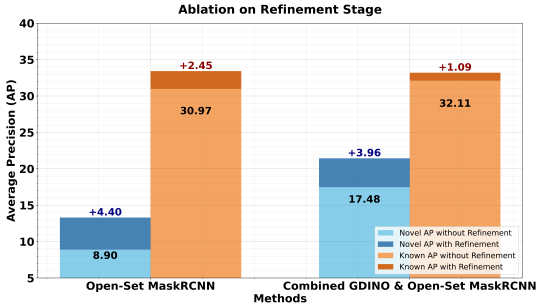


Figure 5. Ablation of our refinement component on the `lvis_val_subset` dataset. The figure reports the Average Precision (AP) for both Novel and Known objects.

a higher Novel AP50 of 54.7%, this comparison is not entirely fair due to the substantial pre-training and training resources it employs (e.g., FILIP, Qformer, GPT-4, and others). While our method uses multiple models during pre-training, it does not use any such model in training (except detector). DetCLIPv3 uses models in both pre-training and training. Due to the zero-shot nature of our method, NOD benchmark in Tab. 1 becomes the main result.

4.4. Comparisons with SOTA Unknown Object Detection Methods

Tab. 3 presents a comparison between our approach and the SOTA in unknown object detection, specifically the UnSniffer model [16]. To align with our evaluation criteria, we retrained UnSniffer on the entire set of COCO classes, excluding its NCut filtering feature for fair comparison. Our results demonstrate a notable improvement over existing methods, such as RNCDL, GDINO, and UnSniffer, in accurately localizing novel and known objects. Our method achieves a leading recall rate of 37.36%, showing an absolute 3.5% gain over the second best method, RNCDL.

Method	Recall (%)	Num TP	Num GT	Tot Pred
GDINO [19]	29.96	13653	45570	223500
UnSniffer [16]	30.83	14048	45570	216424
RNCDL [7]	33.82	15413	45570	216553
Ours*	**37.36**	**17026**	45570	223500

Table 3. Comparison of localization performance of different methods on the `lvis_val_subset` dataset. Recall is measured with an IOU threshold of 0.5. True positives (TP) are those predicted boxes with IOU $>$ 0.5 with any ground truth (GT) box. *Our method with GDINO, CLIP, and Mask-RCNN-V1.

Method	SigLIP	SAM	GDINO	SRM	SAEG	AP (Novel)	AP (Known)	AP (All)
Ours	✗	✓	✓	✓	✓	7.46	**39.44**	9.93
Ours	✓	✗	✓	✓	✓	11.54	34.62	13.32
Ours	✓	✓	✗	✓	✓	11.39	36.81	13.35
Ours	✓	✓	✓	✗	✓	14.32	36.66	16.05
Ours	✓	✓	✓	✓	✗	15.84	37.07	17.48
Ours	✓	✓	✓	✓	✓	**16.12**	37.09	**17.74**

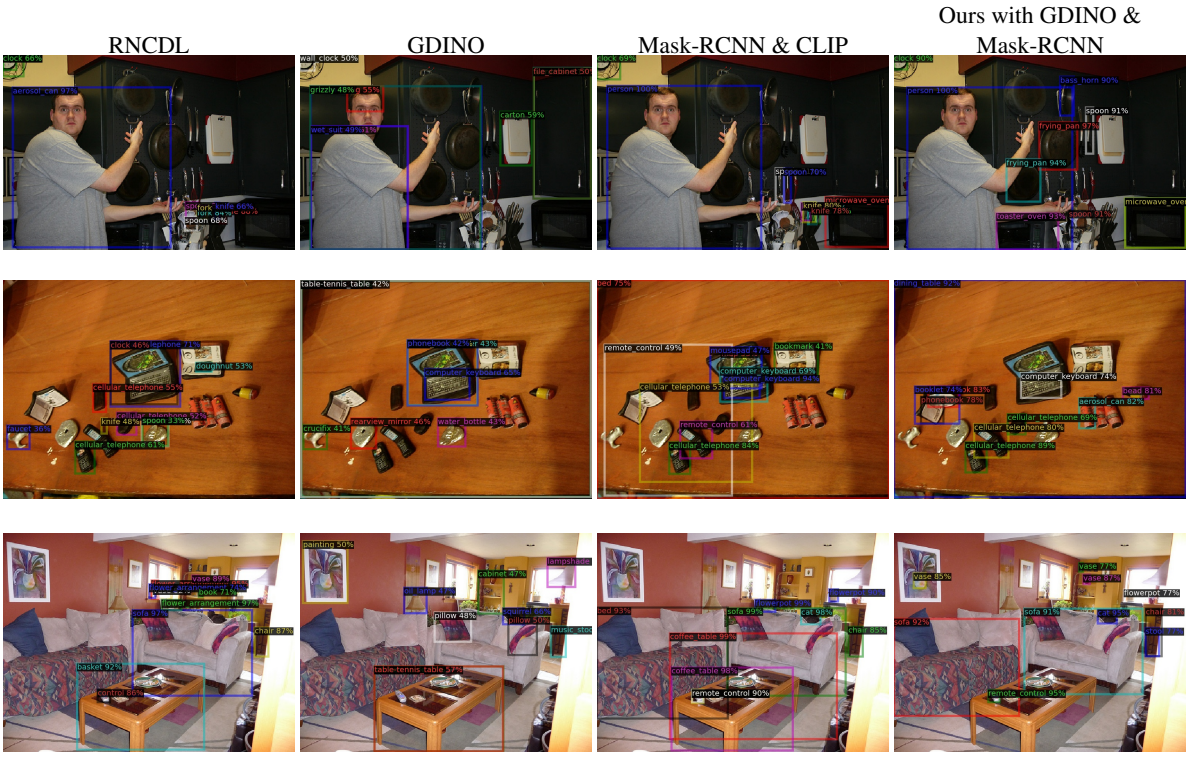
Table 4. Ablation experiments of our method consisting of SigLIP + GDINO + Mask-RCNN-V1 and our proposed components on the `lvis_val` set. Each row represents the performance of the final method minus a particular component indicated by the ✗.

4.5. Ablations

Ablation on Top-N Box Predictions - Grounding DINO vs. Mask-RCNN: We examine the mAP performance of GDINO and Mask-RCNN for both known and novel classes with different numbers of box predictions (Fig. 4). For GDINO, increasing N_{GD} beyond 300 leads to only marginal performance improvements, whereas considering the first 600 boxes leads to only 0.5 mAP increase on novel classes. Considering this alongside GPU memory constraints, we decide to set N_{GD} at 300 for optimal efficiency. In the case of Mask-RCNN, as we can see from Fig. 4, the known AP considering the next 300 boxes is only 1.05 mAP, while the novel AP is mere 0.05 mAP, therefore the top-300 predictions ($N_{\text{KN}} + N_{\text{BG}} = 300$) is optimal, and further predictions introduce more noise and do not significantly capture ground-truth generic objects.

Ablation on Refinement Stage: Our ablation study on the refinement stage (Fig. 5) shows significant improvements due to proposed refinement process. The refinement stage added to open-set Mask RCNN results in a notable 4.4 mAP gain on novel AP. Further, combining GDINO and open-set Mask-RCNN predictions with our refinement stage yields superior performance, resulting in gains of 3.96 mAP in novel, and 1.09 mAP in known classes.

Ablation of Components in our Cooperative Mechanism: Our experiments assess the impact of each component in our method by observing changes in mAP when a component is removed. As indicated in Tab. 4, excluding SigLIP causes the most significant drop in Novel AP (8.66 mAP), suggesting its crucial role in detecting novel



(a) (b) (c) (d)

Figure 6. Comparison of top-10 predictions by different NOD methods. RNCDL (column a) shows mislabelings, such as '`aerosol_can`', '`doughnut`', '`basket`' and '`spoon`' due to post-discovery class assignment issues. GDINO (column b) has several inaccuracies, like '`grizzly`', '`wet_suit`', '`rearview_mirror`', and '`squirrel`'. It also produces a lot of uncertain and low confidence predictions (i.e. score $< 50\%$). Mask-RCNN with CLIP baseline (column c, without our refinement) misses out on objects like '`frying_pan`', and outputs incorrect detections like '`bed`'. Our method (column d, GDINO+Mask-RCNN) accurately detects objects with high confidence and outperforms prior methods in precise localization and recognition of unique objects. Figure best viewed with zoom.

classes. Conversely, this removal slightly improves known AP by 2.35 mAP, highlighting a trade-off between novel and known AP detection. The absence of GDINO reduces novel AP by 4.73 mAP, confirming its importance in identifying novel objects. Similarly, without SAM, there is a 4.58 mAP decrease in novel AP. This implies that even with GDINO and SigLIP, optimal performance on novel classes requires the refinement offered by SAM (Sec. 3.2). Additionally, the performance across both known and novel classes drops significantly without our novel components (SAEG and SRM), demonstrating their effectiveness.

5. Conclusion

In this research, we introduce a novel cooperative mechanism that leverages the complementary strengths of pre-trained foundational models. This mechanism effectively transforms any existing closed-set detector into an open-set detector, addressing the challenges of novel object detection (NOD). The modular design of our approach allows for seamless integration with current open-set detectors, enhancing their performance in detecting both known and novel classes. Our method has consistently outperforms state-of-the-art techniques in various settings, including NOD, unknown object localization, and open-vocabulary detection. While our study primarily focused on bounding box outputs, the underlying components, namely Mask-RCNN and SAM, are inherently capable of generating instance segmentation masks. This aspect suggests that our methodology can be readily extended to instance segmentation tasks, indicating a broader scope of application. Overall, the cooperative mechanism we propose not only addresses current limitations in object detection for NOD but also provides a versatile and effective framework that can be deployed in other applications.

Limitations: Despite the superior performance of our proposed method in NOD, when compared to existing approaches, a primary limitation is the inference speed and potential data leakage. Addressing this limitation (by training efficient custom models from scratch) without compromising its detection capabilities is a key area for future research.

References

[1] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 1, 3, 4

[3] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2722–2730, 2015. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 4

[5] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis, 2022. 2

[6] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery, 2021. 6

[7] Vladimir Fomenko, Ismail Elezi, Deva Ramanan, Laura Leal-Taix'e, and Aljoša Ošep. Learning to discover and detect objects. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 5, 6, 7

[8] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation, 2021. 6

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022. 2, 6

[10] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 1, 2, 5, 6

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1, 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 6

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3, 5

[15] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1, 2

[16] Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3230–3239, 2023. 2, 5, 7

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 6

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 5, 6

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2, 3, 4, 5, 6, 7

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 4

[21] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 6

[22] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training, 2021. 1

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 3, 6

[25] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection, 2022. 6

[26] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *European Conference on Computer Vision*, pages 382–401. Springer, 2022. 2

[27] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020. 1

[28] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 2

[29] Yingjie Wang, Qiuyu Mao, Hanqi Zhu, Jiajun Deng, Yu

Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, pages 1–31, 2023. 1

[30] Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, and Serena Yeung. Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision, 2021. 6

[31] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 6

[32] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection, 2023. 2, 6

[33] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching, 2023. 2, 6

[34] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27391–27401, 2024. 6

[35] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 106–122, Cham, 2022. Springer Nature Switzerland. 2, 6

[36] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions, 2021. 6

[37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 6

[38] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 3

[39] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023. 2

[40] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision, 2022. 6

[41] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1