arXiv:2405.14874v4 [cs.CV] 6 Sep 2024

Open-Vocabulary Object Detectors: Robustness Challenges under Distribution Shifts

Prakash Chandra Chhipa[1,*[0000−0002−6903−7552]], Kanjar De[2[0000−0003−0221−8268]], Meenakshi Subhash Chippa[1,+[0009−0000−2770−6271]], Rajkumar Saini[1[0000−0001−8532−0895]], and Marcus Liwicki[1[0000−0003−4029−6574]]

[1] Luleå Tekniska Universitet, Sweden
{first.middle.last}@ltu.se
+ meechi-2@student.ltu.se
[2] Fraunhofer Heinrich-Hertz-Institut, Berlin, Germany
kanjar.de@hhi.fraunhofer.de
*corresponding author - prakash.chandra.chhipa@ltu.se

Abstract. The challenge of Out-Of-Distribution (OOD) robustness remains a critical hurdle towards deploying deep vision models. Vision-Language Models (VLMs) have recently achieved groundbreaking results. VLM-based open-vocabulary object detection extends the capabilities of traditional object detection frameworks, enabling the recognition and classification of objects beyond predefined categories. Investigating OOD robustness in recent open-vocabulary object detection is essential to increase the trustworthiness of these models. This study presents a comprehensive robustness evaluation of the zero-shot capabilities of three recent open-vocabulary (OV) foundation object detection models: OWL-ViT, YOLO World, and Grounding DINO. Experiments carried out on the robustness benchmarks COCO-O, COCO-DC, and COCO-C encompassing distribution shifts due to information loss, corruption, adversarial attacks, and geometrical deformation, highlighting the challenges of the model's robustness to foster the research in this field. Project webpage: https://prakashchhipa.github.io/projects/ovod_robustness

Keywords: open-vocabulary object detection · foundation model · robustness · distribution shift · zero-shot.

1 Introduction

Recent studies [7] on self-supervised learning have highlighted significant performance impacts under distribution shifts and corruptions, urging enhanced robustness strategies. Similarly, the robustness of foundation models must be examined to understand their resilience against various distribution shifts, corruptions, and adversarial attacks. Foundation AI models, pre-trained on extensive datasets spanning multiple domains, are designed with the primary objective of acquiring a comprehensive understanding of the world and applying their knowledge effectively across a wide range of downstream tasks and applications, thereby facilitating advancements in AI capabilities across multiple fields.

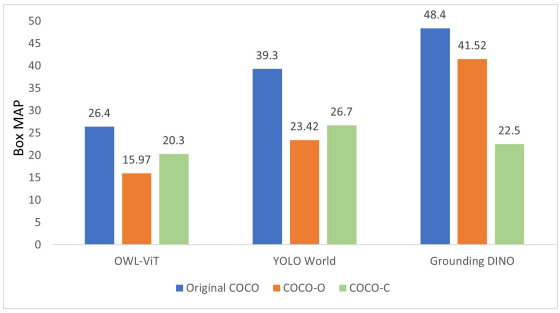

Fig. 1: Zero-shot performance comparison for open vocabulary object detection models, OWL-ViT [36](ECCV'22), YOLO World [6] (CVPR'24), and Grounding DINO [30] (ECCV'24). COCO-O [33] (ICCV'23) represents average results on six subsets, and COCO-C [34] represents average results on fifteen corruptions.

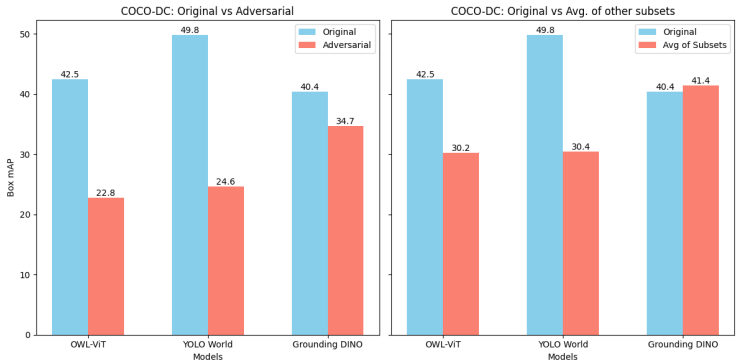

Fig. 2: Zero-shot performance on COCO-DC: (left): comparison for OWL-ViT, YOLO World, and Grounding DINO on COCO-DC robustness performance on original subset and adversarial subset. (right): comparison for these foundation models on COCO-DC robustness performance on original subset and average of all remaining subsets.

The concept of foundation models has been most prominently developed in the fields of natural language processing (NLP) and computer vision (CV). In NLP, foundation models like Generative Pre-trained Transformer (GPT) [1] and Bidirectional Encoder Representations from Transformers (BERT) [10] have revolutionized the field by enabling a range of applications, including text generation, sentiment analysis, question answering, and language translation, without the need for task-specific model architectures.

Models such as Contrastive Language–Image Pre-training (CLIP) [37] and DALL-E [38] demonstrate the ability to understand and generate visual content in response to natural language prompts, showcasing the versatility and creative potential of foundation models. Segment Anything Model (SAM) [24] is a foundation model for image segmentation. The utility of foundation models lies in their ability to leverage their pre-trained knowledge to perform a wide variety of

Fig. 3: Examples from six COCO-O benchmark subsets depicted with predictions by open-vocabulary models: OWL-ViT, YOLO World, and Grounding DINO. The input textual query includes the object categories identified in the labels.

tasks with minimal additional training. This versatility makes them a powerful tool for developing AI applications quickly and efficiently, opening up new possibilities for innovation and research in AI. Open-vocabulary object detection has gained researchers' interest in solving real-world problems. Recent advances in open-vocabulary (OV) object detection in computer vision extend the capabilities of traditional object detection frameworks to recognize and classify objects beyond predefined categories present in their training datasets. Open-vocabulary (OV) models demonstrated zero-shot or few-shot learning capabilities by using vision-language pre-training, making the models to accurately identify and categorize objects they have never encountered during training.

Robustness, defined as a model's ability to maintain consistent performance under varying and unforeseen conditions, has emerged as a critical factor in evaluating the utility of modern machine learning models across diverse applications and data distributions. These conditions include but are not limited to, noisy data, distribution shifts, and adversarial attacks. Robustness involves balancing the trade-off between model accuracy and the ability to generalize well across unforeseen scenarios, challenging the traditional focus on maximizing dataset-specific performance. Robust models contribute to enhanced security, privacy, and user trust, necessitating strategies encompassing data augmentation, regularization techniques, ensemble learning, and more to address challenges like adversarial robustness and long-tail robustness [16].

We explore the robustness of foundation models, with a particular focus on open-vocabulary models. The connection between robustness and the trustworthiness of these models is crucial; robust models inspire greater confidence in their application across various domains, especially when leveraging zero-shot and few-shot learning capabilities. By bringing attention to the topic of robustness of open-vocabulary models, we inspire to improve trustworthiness by investigating the out-of-distribution performance on zero-shot evaluation.

To the best of our knowledge, we are the first study to provide comparative analysis of recent state-of-the-art OV object detectors: OWL-ViT [36], YOLO-World [6], and Grounding DINO [30] for open vocabulary object detec-

tion through the lenses of robustness. OWL-ViT leverages the Vision Transformer (ViT) for transfer learning, YOLO-World builds on the efficient and practical CNN architecture, and Grounding DINO integrates the Swin Transformer [31] with a novel grounded pre-training strategy.

We employ a zero-shot evaluation approach (Refer Figure 3), where models are tested on the out-of-distribution benchmarks such as COCO-O [33] and COCO-C [34], without any additional training on its specific subsets. COCO-O offers six subsets with decreasing details of objects in terms of shape, color, and textures, whereas COCO-C comprises fifteen subsets corresponding to corruptions from [18]. This evaluation investigates the models' ability to generalize from learned representations to unseen categories and conditions. The notable observation suggests that all three open-vocabulary foundation model-based object detectors, when subjected to degradation of image quality and distribution shift, exhibit significant deviations in performance. This indicates an inherent relationship between OV object detectors and the quality of data. Therefore, this work draws attention to the research community and motivates further research towards improving robustness (refer Figures 1 and 2).

2 Related Work

AlexNet [25] revolutionized computer vision by applying deep learning, which was further advanced by Fast R-CNN [14] and Faster R-CNN [40], enhancing proposal classification and generation. RetinaNet's [29] Focal Loss addressed class imbalance, while YOLO [39] achieved real-time detection through a unified regression framework. The Vision Transformer (ViT) [11] innovated by applying transformers to image patches, demonstrating their potential beyond NLP, and the Swin Transformer [31] introduced a hierarchical structure for efficient image processing, setting new standards in vision benchmarks.

Carion et al. [4] developed the DEtection TRansformer (DETR), redefining object detection as a set prediction problem. Utilizing a transformer architecture with an encoder-decoder and a global set-based loss for unique prediction via bipartite matching, DETR streamlines the detection process, removing the necessity for components like non-maximum suppression. However, the detection capability of such models is confined to their trained categories, limiting their effectiveness on unseen objects. To address these constraints, the introduction of foundation models [27, 37] has been proposed.

Foundation AI models like GPT [1] and BERT [10] in NLP, and CLIP [37], DALL-E [38], and SAM [24] in computer vision, demonstrate the power of large-scale pre-training across data-rich domains. These models have revolutionized tasks like text generation, sentiment analysis, and visual content understanding. Their vast parameter space and training breadth enable extensive adaptability, facilitating the rapid development of specialized AI applications and propelling forward AI research.

Radford et al.'s CLIP [37] redefines visual learning with natural language, pre-trained on 400 million web-sourced image-text pairs. This method enables

understanding of various visual concepts and supports zero-shot transfer to tasks without specific training. Tested on over 30 benchmarks, including OCR, action recognition, and geo-localization; CLIP's adaptability significantly expands visual model applications.

The Grounded Language-Image Pre-training (GLIP) [27] advances visual learning by fusing language supervision, setting new zero-shot and few-shot learning standards. It leverages 27 million annotated and web-sourced data points to enhance object recognition. Open-vocabulary detection, by leveraging NLP, allows models to recognize new objects from descriptions, improving adaptability and robustness in changing environments.

To the best of our knowledge, Zareian et al. [46] first open-vocabulary object detection using image-caption pairs for detecting unannotated objects, outperforming zero-shot and weakly supervised approaches by leveraging visual-semantic spaces from captions. Chen et al. [5] developed MEDet, enhancing Open-Vocabulary Object Detection by aligning vision-language at the proposal level and balancing predictions between known and novel categories, showing top results on MS COCO and LVIS. Bravo et al. [3] proposed a method for open-vocabulary detection that uses localized vision-language matching to improve alignment of visual and linguistic representations, aiming to expand detection vocabularies. Zhao et al. [48] demonstrated how combining Vision and Language models with unlabeled data can improve object detection, highlighting the synergy between pre-trained models and large, unlabeled datasets. Zang et al. [45] introduced OV-DETR, extending DETR for open-vocabulary detection using conditional matching with CLIP-generated embeddings, achieving significant advances on LVIS and COCO. RegionCLIP [49] tackled domain shift in open-vocabulary detection by aligning regional visual representations with text, surpassing existing methods on COCO and LVIS for novel and zero-shot categories. Du et al. [12] presented DetPro, a method for learning continuous prompt representations, enhancing the detection of novel classes through innovative proposal handling and context grading on the LVIS dataset. Feng et al. [13] created PromptDet, combining CNNs and transformers for object detection, leveraging spatial and global context for improved accuracy and robustness in detection tasks.

Kaul et al. [21] present a multi-modal OVOD method, surpassing the LVIS benchmark by fusing text and visual classifiers with large language models. Cho et al. [8] innovate in novel object detection using PCL for captions, enhancing LVIS performance via image captioning model distillation. Arandjelovic et al. [2] show that combining semantic segmentation with detection improves accuracy in complex scenes. Minderer et al. [35] explore the benefits of self-training on detection models with large image-text datasets. Shi et al. [41] develop an OV detection framework based on scene graphs, validated by comprehensive tests. Zhao et al. [47] introduce SAS-Det, a method tackling noisy pseudo labels for improved detection, achieving high COCO and LVIS scores. Kim et al. [22] propose CFM-ViT, a contrastive learning approach for OV detection that excels on LVIS. Kuo et al. [26] demonstrate efficient OV detection with F-VLM by

training only the detector head. Kim et al. [23] present RO-ViT, a pretraining strategy enhancing OV detection alignment, leading to top LVIS and COCO results. Finally, Wang et al. [43] reveal OADP, an approach for transferring knowledge to OV detectors, outdoing current methods on MS-COCO, and Yao et al. [44] introduce DetCLIPv2, a scalable OVD training framework that sets a new zero-shot AP record on LVIS.

Robustness is crucial for assessing the resilience and reliability of machine learning models, emphasizing their ability to perform consistently under variable and unexpected conditions, such as noisy data, distribution shifts, and adversarial attacks. It shows the importance of a model's capacity to generalize beyond its training, ensuring dependable predictions against non-standard inputs. This necessitates a delicate balance between accuracy and generalization, moving away from solely focusing on dataset-specific performance to prevent overfitting. Addressing robustness requires strategies like data augmentation, regularization, and ensemble learning to combat adversarial threats and variability, enhancing model security, privacy, and trustworthiness [16]. These measures are essential for the practical application of AI in diverse, changing real-world scenarios, aiming to create accurate, secure, and adaptable systems. To the best of our knowledge, Hendrycks et al. [17] proposed one of the earliest datasets Imagenet-C to benchmark robustness. Over the years, different versions of datasets have been derived from the original Imagenet dataset to study the topic of robustness for image classification models. Some of the notable datasets are Imagenet-A [19], Imagenet-R [15], Imagenet-CD [9], Imagenet-E [28], Imagenet-X [20] and Imagenet-Sketch [42]. However, very little work is available in the literature to study the impact of distribution shifts on object detection models.

3 Out-of-Distribution benchmarks

In this section, We briefly discuss two OOD benchmarks for the robustness evaluation, namely COCO-O [33], COCO-DC [32] and COCO-C [34].

3.1 COCO-O

Recent robustness benchmark COCO-O [33] dataset dedicated to pose the challenge of object detection under natural distribution shifts, serving as a comprehensive benchmark for assessing detector robustness beyond the typical constraints of existing datasets. COCO-O encompasses a range of challenges inherent to object detection, including occlusion, blurring, variations in pose, deformation, illumination differences, and the detection of small-sized objects. COCO-O comprises 6,782 images collected online across six distinct subsets: weather, painting, handmake, cartoon, tattoo, and sketch, arranged in descending order based on the level of detail present within the objects they contain. This organization reflects varying degrees of abstraction across the objects within each domain.

3.2 COCO-DC

The COCO-DC [32] object detection robustness benchmark is recently curated from the COCO 2017 validation set, comprising 1,127 images that distinctly separate foreground objects from their backgrounds. This dataset is designed to evaluate the robustness of object detection and classification models under various background conditions. The COCO-DC benchmark features four subsets: Adversarial, BLIP-2 Caption, Color, and Texture. The Adversarial subset includes images with adversarial background changes crafted to challenge the models' robustness. The BLIP-2 Caption subset utilizes the BLIP-2 model to generate captions for the images, providing a different context for evaluation. The Color subset features images with altered background colors to assess the models' performance under color variations. The Texture subset consists of images with modified background textures to test the models' resilience to texture changes. This benchmark allows for a comprehensive analysis of model performance across diverse and challenging scenarios, highlighting the strengths and weaknesses of contemporary vision-based models in handling object-to-background context variations.

3.3 COCO-C

COCO-C [34] dataset introduces 15 types of image corruptions, each with five levels of severity, covering a broad range of corruption types sorted into four groups: noise, blur, digital, and weather. This comprehensive approach enables a nuanced assessment of model robustness across different distortion types and severity levels, which are not part of the original training regime. The datasets are not intended for training data augmentation but rather to measure a model's robustness against unseen corruptions, thus helping to identify areas for improvement in object detection models.

4 Open-Vocabulary Object detectors Models

In this section, we describe the three open-vocabulary foundation models: OWL-ViT, YOLO World, and Grounding DINO.

4.1 OWL-ViT

OWL-ViT [36] method introduces an efficient and effective solution for open-vocabulary object detection by leveraging the Vision Transformer (ViT) architecture with minimal modifications and a comprehensive strategy for transferring image-text pre-training to the task of object detection. At its core, OWL-ViT employs a standard ViT for image encoding, which, during the transfer phase to detection, is slightly altered by removing the final token pooling layer and adding lightweight object classification and box prediction heads directly to the

output tokens. This design choice allows for the direct prediction of object instances without the need for additional complex mechanisms. The methodology progresses through integrating text embeddings derived from a pretrained language model, facilitating open-vocabulary classification capabilities. This approach strengthens the model to identify a diverse spectrum of object categories beyond those encountered during its initial training phase. Through the end-to-end fine-tuning of both visual and linguistic components on conventional object detection datasets, OWL-ViT showcases exceptional efficacy across various evaluation benchmarks. This performance is attributed to the strategic utilization of extensive image-text corpora for pre-training, succeeded by meticulous fine-tuning processes. Consequently, OWL-ViT establishes new benchmarks in the domains of zero-shot, text-conditioned, and one-shot, image-conditioned object detection tasks.

4.2 YOLO-World

YOLO-World [6] introduces a novel open-vocabulary object detection framework that significantly enhances the conventional YOLO detection model with the capacity for open-vocabulary detection, achieving real-time efficiency and high accuracy across diverse benchmarks. YOLO-World incorporates the Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) and a distinct region-text contrastive loss. These components work in tandem to ensure a robust visual-semantic alignment between image features and textual embeddings. This strategic integration strengthens the model's ability to adeptly navigate the complex interplay between visual and linguistic information, significantly enhancing its open-vocabulary detection capabilities while maintaining real-time processing speeds. This approach leverages large-scale datasets for pre-training, effectively combining detection, grounding, and image-text data into a unified learning framework. As a result, YOLO-World not only extends the capabilities of the YOLO architecture to recognize a broader array of object categories in a zero-shot manner but also does so with remarkable inference speed and deployment efficiency.

4.3 Grounding DINO

Grounding DINO [30] a novel approach to open-set object detection by integrating the strengths of the Transformer-based detector DINO with grounded pre-training techniques. This method allows for detecting arbitrary objects based on human input, such as category names or referring expressions, by effectively fusing language and vision modalities. Grounding DINO partitions a closed-set detector into three phases—feature enhancement, language-guided query selection, and cross-modality decoder—to achieve a tightly integrated fusion of language and vision. Unlike prior works that evaluate open-set object detection primarily on novel categories, Grounding DINO also extends evaluations to referring expression comprehension (REC), enabling the model to understand objects specified with attributes.

5 Experiments and Results

The list of the object categories present in the input image is used as a text query for zero-shot evaluation of the open vocabulary object detection models (refer to Figure 3). This work evaluates OWL-ViT, YOLO-World, and Grounding DINO models on all six OOD subsets of COCO-O, 5 subsets of COCO-DC [32], and on fifteen corruption subsets (at severity level 1) of COCO-C benchmark [15].

5.1 Evaluation Method

This work evaluates the robustness of OWL-ViT, YOLO World, and Grounding DINO models for their zero-shot capability based on their performance on OOD benchmarks COCO-O, COCO-DC, and COCO-C, as shown in Figure 4.

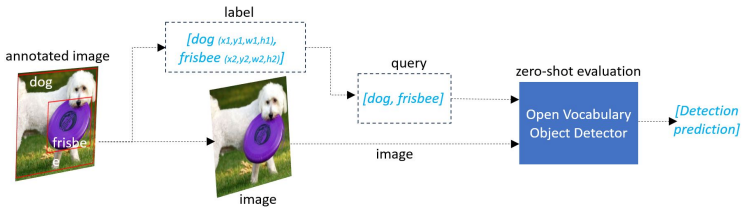

Fig. 4: Zero-shot evaluation process of open vocabulary object detector models.

5.2 Metrics

AP (Average Precision) measures the precision of a model at different confidence thresholds, summarizing its detection accuracy for a specific class. It is the area under the precision-recall curve.

mAP@IoU=0.5 (mean Average Precision at Intersection over Union of 0.5) averages the AP values for all classes at an IoU threshold of 0.5, meaning detections are correct if the predicted and ground truth boxes overlap by at least 50%. Box mAP@IoU=0.5:0.95 evaluates the model across IoU thresholds from 0.5 to 0.95, in steps of 0.05. It calculates the AP at each threshold and averages these values, providing a rigorous assessment of detection precision at varying overlap levels.

The effective robustness metric $ER(f)$ for a detector f is computed as given in Eq. 1. mAP_{ood} is the mAP on COCO-O dataset, and mAP_{id} is the mAP on original COCO dataset. mAP_{id} is multiplied by a factor β in the equation. The value of β is fixed to 0.45, adopted from [33].

$$ER(f) = mAP_{ood}(f) - \beta \times mAP_{id}(f) \tag{1}$$

5.3 Discussions

Detailed results on six subsets of COCO-O benchmarks are described in Table 1 and 2, results on subsets of COCO-DC benchmark is in Table 3, and for 15 corruptions subsets of COCO-C benchmarks, results are in Table 4, 5, 7, and 6.

Table 1: Comparison of Box mAP (@IoU=0.50:0.95) for different detectors in original COCO and COCO-O datasets. *referred from Grounding DINO [30].

Box mAP@IoU=0.50:0.95									
	Venue	COCO mAP	COCO-O mAP						
			Sketch	Weather	Cartoon	Painting	Tattoo	Handmake	Average
OWL-ViT	ECCV'22	26.40	14.60	19.50	17.20	24.50	7.70	12.30	15.97
YOLO World	CVPR'24	39.30	15.00	37.90	18.40	36.00	10.10	23.10	23.42
Grounding DINO	ECCV'24	48.40*	44.90	33.70	47.50	42.30	41.10	39.60	41.52

Table 2: Comparison of mAP (@IoU=0.50) for different detectors in original COCO and COCO-O datasets.

mAP@IoU=0.50									
	Venue	COCO mAP	COCO-O mAP						
			Sketch	Weather	Cartoon	Painting	Tattoo	Handmake	Average
OWL-ViT	ECCV'22	42.90	18.90	31.00	23.50	33.30	9.80	16.10	22.10
YOLO World	CVPR'24	51.20	17.30	46.00	22.50	43.80	12.40	27.40	28.23
Grounding DINO	ECCV'24	-	53.00	40.90	55.90	49.50	50.90	45.20	49.23

COCO-O: Following the results in Table 1 and 2, OWL-ViT on the original COCO dataset stands at 26.40 and 42.90 for IoU=0.50:0.95 and IoU=0.50, respectively. When subjected to the COCO-O dataset, the average performance drops to 15.97 and 22.10 across the same IoU thresholds, indicating a significant decrease in robustness under out-of-distribution conditions. Similarly, YOLO World, which achieves a higher baseline mAP of 39.30 and 51.20 on the COCO dataset for IoU=0.50:0.95 and IoU=0.50, respectively, shows a reduced average mAP of 23.42 and 28.23 on COCO-O. This suggests a resilience to out-of-distribution data compared to OWL-ViT, though the performance still notably decreases. Grounding DINO demonstrates the most remarkable performance, with a mAP of 48.40 on COCO for IoU=0.50:0.95 and a consistent lack of data for IoU=0.50. On COCO-O, it achieves an average mAP of 41.52 and 49.23 across the respective IoU thresholds. Grounding DINO exhibits the least performance drop among the three, indicating robustness to out-of-distribution scenarios, as indicated the trend of effective robustness of models in Fig. 5.

COCO-DC: The performance of three open-vocabulary object detectors—OWL-ViT, YOLO World, and Grounding DINO—on the COCO-DC dataset subsets is summarized in Table 3. All three models exhibit high performance on the Original subset, with YOLO World achieving the highest Box mAP (49.8), followed by OWL-ViT (42.5) and Grounding DINO (40.4). However, the Adversarial subset causes a significant drop for all models. Grounding DINO demonstrates the highest robustness (34.7), while YOLO World and OWL-ViT drop to 24.6 and 22.8, respectively. On the BLIP-2 Caption subset, Grounding DINO (42.6) and YOLO World (39.5) perform well, but OWL-ViT lags (34.2).

The Color and Texture subsets challenge all models, with pronounced performance drops. Grounding DINO maintains the highest scores (44.4 and 43.9), showing superior adaptability. YOLO World scores 29.3 for Color and 28.4 for Texture, while OWL-ViT scores 33.4 and 30.6, respectively. This suggests Grounding DINO's training strategy allows better generalization across visual distortions. Overall, while Grounding DINO exhibits the highest resilience, all models show vulnerabilities under distribution shifts, emphasizing the need for

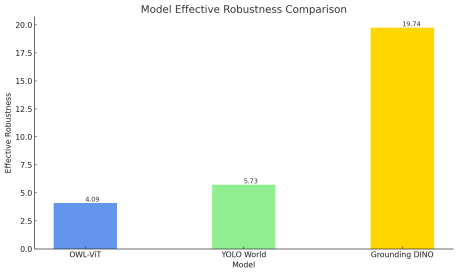

Fig. 5: Comparisons of effective robustness for detectors based on their performance on original COCO and COCO-O datasets.

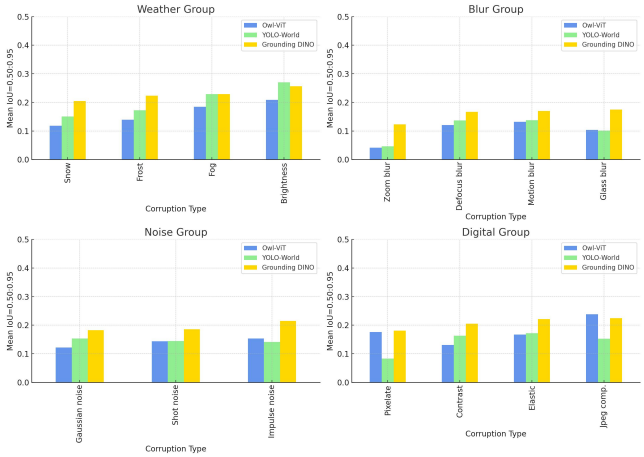

Fig. 6: Comparisons of performance of the models (Owl-ViT, YOLO-World, and Grounding DINO) on four corruptions group of COCO-C.

diverse and challenging training scenarios to enhance robustness in real-world applications.

COCO-C: In the robustness evaluation conducted on the COCO-C dataset, three models—Owl-ViT, YOLO-World, and Grounding DINO were assessed across fifteen different corruption types grouped into Weather, Blur, Noise, and Digital categories (detailed results in Table 4, 5, 6, and 7).

The results reveal a trend of decreasing performance with increasing severity of corruption across all models and corruption types, highlighted by notable performance drops such as Owl-ViT's decrease from 40.0 to 12.6 IoU in Pixelate corruption from severity 1 to 5.

YOLO-World generally displayed superior resilience in lower severity levels, particularly in handling weather-related and blur corruptions, where it outperformed Owl-ViT, managing a 37.5 IoU at severity 1 for Frost compared to Owl-ViT's 32.4. Conversely, Grounding DINO, while slightly underperforming in higher intersection over union (IoU) metrics, showed comparable or better performance in detecting objects at the basic IoU=0.50 level across many cor-

Table 3: Comparison of Box mAP (@IoU=0.50:0.95) for different detectors in the COCO-DC dataset subsets.

Box mAP@IoU=0.50:0.95					
	Original	**Adversarial**	**BLIP**	**Color**	**Texture**
OWL-ViT	42.5	22.8	34.2	33.4	30.6
YOLO World	49.8	24.6	39.5	29.3	28.4
Grounding DINO	40.4	34.7	42.6	44.4	43.9

Table 4: Robustness evaluation on Weather group: Snow, Frost, Fog, and Brightness corruption subsets in the COCO-C dataset.

Corruption	Severity	Owl-ViT		YOLO-World		Grounding DINO	
		IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50
Snow	1	18.0	29.9	25.2	32.7	24.7	34.4
Snow	2	12.8	20.6	14.4	19.3	21.2	30.2
Snow	3	11.9	19.8	14.0	18.8	20.4	28.8
Snow	4	8.3	13.7	10.9	14.8	18.6	27.3
Snow	5	8.2	13.4	10.8	14.6	17.3	25.5
Frost	1	19.8	32.4	28.9	37.5	24.6	33.9
Frost	2	15.5	25.2	17.8	23.6	22.6	31.9
Frost	3	12.2	19.9	14.6	19.4	20.5	29.2
Frost	4	12.1	19.4	14.0	18.5	20.3	28.8
Frost	5	10.1	16.4	11.7	15.5	18.9	27.5
Fog	1	20.9	33.7	29.8	38.4	23.9	32.7
Fog	2	19.6	31.6	22.6	29.8	23.4	32.0
Fog	3	18.6	29.8	21.5	28.3	23.0	31.4
Fog	4	17.6	28.2	21.3	28.1	22.7	31.1
Fog	5	15.4	24.5	19.6	25.8	21.9	30.3
Brightness	1	23.3	37.8	33.5	43.5	26.1	35.5
Brightness	2	22.4	36.2	27.0	35.6	26.1	35.6
Brightness	3	21.2	34.2	26.2	34.6	25.9	35.5
Brightness	4	19.6	31.4	25.0	33.1	25.5	35.0
Brightness	5	17.9	28.5	23.4	31.0	24.9	34.5
Original COCO	-	26.40	42.90	39.30	51.20	48.40	-

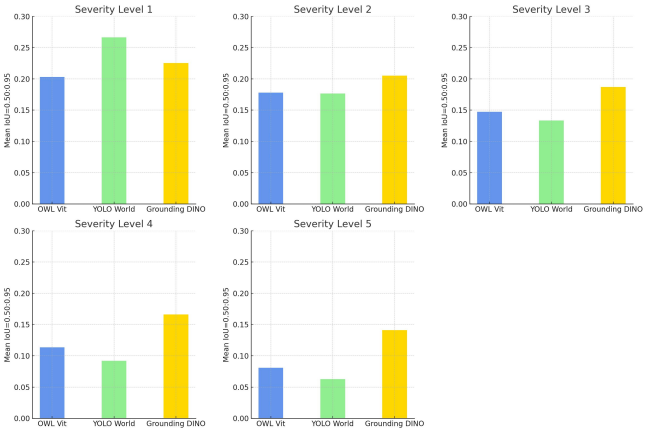

Fig. 7: Comparisons of performance of the models (Owl-ViT, YOLO-World, and Grounding DINO) on across severity level of underlying corruptions in COCO-C.

ruption scenarios, such as achieving an IoU of 35.5 in high-severity Brightness corruption versus Owl-ViT's 28.5. As trend shown in Figure 6, across all groups suggests that while Grounding DINO often excels in handling blur-induced distortions, Owl-ViT generally lags behind, especially as the complexity of corruptions increases. A meta-comparison (Fig. 1 reveals varying degrees of resilience among the models. Grounding DINO stands out for its robustness, maintaining

Table 5: Robustness evaluation on Blur group: Zoom blur, Defocus blur, Motion blur, and Glass blur corruption subsets in the COCO-C dataset.

Corruption	Severity	Owl-ViT		YOLO-World		Grounding DINO	
		IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50
Zoom blur	1	8.3	15.8	12.6	20.8	13.3	22.6
Zoom blur	2	5.1	10.5	4.7	8.8	10.0	18.6
Zoom blur	3	3.4	7.4	3.0	6.0	8.1	16.1
Zoom blur	4	2.2	5.1	1.8	3.9	6.4	13.5
Zoom blur	5	1.6	3.9	1.2	2.7	5.4	11.9
Defocus blur	1	17.9	28.1	28.0	36.3	20.7	28.2
Defocus blur	2	15.5	24.1	19.4	26.0	18.8	26.2
Defocus blur	3	11.9	18.4	14.0	19.2	16.7	23.9
Defocus blur	4	9.1	14.1	9.1	12.7	14.2	20.5
Defocus blur	5	6.9	10.8	5.1	7.2	12.8	18.7
Motion blur	1	19.4	31.8	27.9	37.3	21.8	30.1
Motion blur	2	16.2	26.7	17.3	24.5	19.5	26.2
Motion blur	3	12.0	20.3	11.0	16.0	16.0	22.0
Motion blur	4	7.8	13.3	5.2	7.9	12.5	19.1
Motion blur	5	5.5	9.2	2.8	4.4	10.6	18.2
Glass blur	1	17.3	26.9	25.7	33.7	21.1	28.5
Glass blur	2	14.2	22.0	15.0	20.1	15.4	19.8
Glass blur	3	8.3	12.9	4.5	6.3	12.6	17.6
Glass blur	4	6.8	10.5	3.1	4.3	11.2	15.7
Glass blur	5	5.1	7.9	1.9	2.6	9.3	13.5
Original COCO	-	26.40	42.90	39.30	51.20	48.40	-

Table 6: Robustness evaluation on Noise group: Gaussian noise, Shot noise, and Impulse noise corruption subsets in the COCO-C dataset.

Corruption	Severity	Owl-ViT		YOLO-World		Grounding DINO	
		IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50
Gaussian noise	1	22.4	36.8	28.2	36.8	21.8	29.6
Gaussian noise	2	19.3	32.0	23.4	31.0	19.3	26.2
Gaussian noise	3	14.9	24.7	16.5	22.0	19.4	27.3
Gaussian noise	4	9.9	16.5	8.5	11.4	17.7	25.9
Gaussian noise	5	4.6	7.5	2.1	2.8	12.7	19.1
Shot noise	1	22.7	37.4	28.4	37.1	22.2	29.9
Shot noise	2	19.3	31.9	18.2	24.5	20.2	27.5
Shot noise	3	14.8	24.4	12.5	17.0	18.5	24.7
Shot noise	4	9.4	15.4	4.7	6.5	15.2	21.0
Shot noise	5	5.5	8.9	1.7	2.4	12.1	18.0
Impulse noise	1	19.6	32.2	25.8	33.8	22.7	30.9
Impulse noise	2	17.1	28.0	15.6	20.9	21.4	29.2
Impulse noise	3	14.7	24.3	11.7	15.9	20.4	28.6
Impulse noise	4	9.9	16.5	4.5	6.1	18.4	26.7
Impulse noise	5	5.1	8.5	0.9	1.3	13.4	20.1
Original COCO	-	26.40	42.90	39.30	51.20	48.40	-

closer performance levels between the original and out-of-distribution benchmarks. YOLO-World shows moderate resilience, with a noticeable but smaller performance drop compared to OWL-ViT, which experiences the most significant decrease in mAP when transitioning from COCO to OOD benchmarks.

Increased severity levels: Figure 7 illustrates the decline in performance across five severity levels for the Owl-ViT, YOLO-World, and Grounding DINO models. YOLO-World consistently outperforms the other models across all severity levels, maintaining a higher mean IoU, particularly at severity level 1 where it achieves a performance peak notably higher than its counterparts. As severity increases, all models demonstrate a downward trend, with Owl-ViT having a substantial drop, especially notable between severity levels 1 and 4. Grounding DINO, while not leading at lower severities, shows a more gradual decline, suggesting a degree of robustness in more challenging conditions, as its performance at severity level 5 remains competitive with YOLO-World's. The consistency of these results with the earlier detailed performance metrics across various corruption types validates the trend that model robustness significantly diminishes with increased corruption severity. This emphasizes the importance of robustness evaluations across different levels of corruption severity to assess the reliability

Table 7: Robustness evaluation on Digital group: Pixelate, Contrast, Elastic Transform, and Jpeg compression corruption subsets in the COCO-C dataset.

Corruption	Severity	Owl-ViT		YOLO-World		Grounding DINO	
		IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50	IoU=0.50:0.95	IoU=0.50
Pixelate	1	24.8	40.0	23.4	30.2	23.3	31.1
Pixelate	2	24.3	39.1	13.5	17.8	21.5	28.4
Pixelate	3	17.7	28.0	4.9	6.5	17.5	23.3
Pixelate	4	12.8	19.9	1.9	2.7	13.8	18.4
Pixelate	5	8.5	12.6	0.6	0.7	9.9	13.5
Contrast	1	21.1	33.9	29.7	38.3	23.7	32.3
Contrast	2	19.5	30.9	21.8	28.7	23.0	31.3
Contrast	3	16.3	25.6	18.4	24.1	22.0	30.0
Contrast	4	9.2	14.1	9.4	12.5	16.0	24.0
Contrast	5	2.3	3.5	1.6	2.0	11.0	16.1
Elastic	1	21.6	35.3	27.9	36.7	24.0	33.5
Elastic	2	19.4	32.3	19.1	26.2	22.8	32.3
Elastic	3	16.8	28.2	14.8	20.7	20.5	30.1
Elastic	4	14.4	24.4	11.7	16.5	19.3	28.2
Elastic	5	11.6	20.0	8.2	11.8	17.2	25.5
Jpeg comp.	1	27.3	44.6	24.9	32.5	24.3	33.6
Jpeg comp.	2	26.9	44.3	15.2	20.7	22.4	30.6
Jpeg comp.	3	26.5	44.1	12.4	16.7	19.0	26.2
Jpeg comp.	4	21.4	36.5	6.8	9.3	17.3	24.3
Jpeg comp.	5	13.0	22.4	2.6	3.5	14.3	20.3
Original COCO	-	26.40	42.90	39.30	51.20	48.40	-

of models in real-world scenarios. These findings emphasize the need for further research on robust models capable of maintaining performance under varying degrees of corruption. The consistent performance drop across models points to an essential area for future research: enhancing object detection models' adaptability and resilience against varied and newer visual concept.

6 Conclusion

To the best of our knowledge, we are making one of the earliest attempts to understand zero-shot evaluation on open-vocabulary foundation models through the perspective of robustness under distribution shifts. Through extensive analysis of three recent open-vocabulary foundation object detection models on three public benchmarks, we show that object detection under conditions of out-of-distribution (OOD) shifts poses significant challenges regarding performance deviation, advocating increased focus and investigation by the research community. Using vision-language models combined with effective, prompt engineering can be the future direction for developing more robust open-vocabulary object detectors. Enhancing robustness against various distribution shifts increases the trustworthiness of open-vocabulary object detection models, potentially leading to their adoption across diverse applications.

7 Acknowledgment

The authors thank Sumit Rakesh, Luleå University of Technology, for his support with the Lotty Bruzelius cluster. We also thank the National Supercomputer Centre at Linköping University for the Berzelius supercomputing, supported by the Knut and Alice Wallenberg Foundation.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2, 4
2. Arandjelović, R., Andonian, A., Mensch, A., Hénaff, O.J., Alayrac, J.B., Zisserman, A.: Three ways to improve feature alignment for open vocabulary detection. arXiv preprint arXiv:2303.13518 (2023) 5
3. Bravo, M.A., Mittal, S., Brox, T.: Localized vision-language matching for open-vocabulary object detection. In: DAGM German Conference on Pattern Recognition. pp. 393–408. Springer (2022) 5
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
5. Chen, P., Sheng, K., Zhang, M., Lin, M., Shen, Y., Lin, S., Ren, B., Li, K.: Open vocabulary object detection with proposal mining and prediction equalization. arXiv preprint arXiv:2206.11134 (2022) 5
6. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. Conference on Computer Vision and Pattern Recognizion (CVPR) (2024) 2, 3, 8
7. Chhipa, P.C., Holmgren, J.R., De, K., Saini, R., Liwicki, M.: Can self-supervised representation learning methodswithstand distribution shifts and corruptions? In: 2023 IEEE/CVF International Conference on Computer Vision Workshops. Workshop and Challenges for Out-of-Distribution Generalization in Computer Vision). pp. 4467–4476 (2023) 1
8. Cho, H.C., Jhoo, W.Y., Kang, W., Roh, B.: Open-vocabulary object detection using pseudo caption labels. arXiv preprint arXiv:2303.13040 (2023) 5
9. De, K., Pedersen, M.: Impact of colour on robustness of deep neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21–30 (2021) 6
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2, 4
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
12. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14084–14093 (2022) 5
13. 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: European Conference on Computer Vision. pp. 701–717. Springer (2022) 5
14. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015) 4
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021) 6, 9

16. Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J.: Unsolved problems in ml safety. arXiv preprint arXiv:2109.13916 (2021) 3, 6
17. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019) 6
18. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and surface variations. arXiv preprint arXiv:1807.01697 (2018) 4
19. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021) 6
20. Idrissi, B.Y., Bouchacourt, D., Balestriero, R., Evtimov, I., Hazirbas, C., Ballas, N., Vincent, P., Drozdzal, M., Lopez-Paz, D., Ibrahim, M.: Imagenet-x: Understanding model mistakes with factor of variation annotations. arXiv preprint arXiv:2211.01866 (2022) 6
21. Kaul, P., Xie, W., Zisserman, A.: Multi-modal classifiers for open-vocabulary object detection. In: International Conference on Machine Learning. pp. 15946–15969. PMLR (2023) 5
22. Kim, D., Angelova, A., Kuo, W.: Contrastive feature masking open-vocabulary vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15602–15612 (2023) 5
23. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11144–11154 (2023) 6
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 2, 4
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012) 4
26. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022) 5
27. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022) 4, 5
28. Li, X., Chen, Y., Zhu, Y., Wang, S., Zhang, R., Xue, H.: Imagenet-e: Benchmarking neural network robustness via attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20371–20381 (2023) 6
29. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 4
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. European Conference on Computer Vision (ECCV) (2024) 2, 3, 8, 10
31. 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: Proceedings

of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 4
32. Malik, H.S., Huzaifa, M., Naseer, M., Khan, S., Khan, F.S.: Objectcompose: Evaluating resilience of vision-based models on object-to-background compositional changes. arXiv preprint arXiv:2403.04701 (2024) 6, 7, 9
33. Mao, X., Chen, Y., Zhu, Y., Chen, D., Su, H., Zhang, R., Xue, H.: Coco-o: A benchmark for object detectors under natural distribution shifts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6339–6350 (2023) 2, 4, 6, 9
34. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019) 2, 4, 6, 7
35. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024) 5
36. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision (ECCV). pp. 728–755. Springer (2022) 2, 3, 7
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 4
38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021) 2, 4
39. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 4
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 4
41. Shi, H., Hayat, M., Cai, J.: Open-vocabulary object detection via scene graph discovery. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4012–4021 (2023) 5
42. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. pp. 10506–10518 (2019) 6
43. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2023) 6
44. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023) 6
45. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. pp. 106–122. Springer (2022) 5

46. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021) 5
47. Zhao, S., Schulter, S., Zhao, L., Zhang, Z., Suh, Y., Chandraker, M., Metaxas, D.N., et al.: Improving pseudo labels for open-vocabulary object detection. arXiv preprint arXiv:2308.06412 (2023) 5
48. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: European Conference on Computer Vision. pp. 159–175. Springer (2022) 5
49. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022) 5