

YOLO-UniOW: Efficient Universal Open-World Object Detection

Lihao Liu^{1,*}, Juexiao Feng^{1,*}, Hui Chen¹, Ao Wang¹, Lin Song², Jungong Han¹, Guiguang Ding^{1,✉}
¹ Tsinghua University ² Tencent ARC Lab

Abstract

Traditional object detection models are constrained by the limitations of closed-set datasets, detecting only categories encountered during training. While multimodal models have extended category recognition by aligning text and image modalities, they introduce significant inference overhead due to cross-modality fusion and still remain restricted by predefined vocabulary, leaving them ineffective at handling unknown objects in open-world scenarios. In this work, we introduce Universal Open-World Object Detection (Uni-OWD), a new paradigm that unifies open-vocabulary and open-world object detection tasks. To address the challenges of this setting, we propose YOLO-UniOW, a novel model that advances the boundaries of efficiency, versatility, and performance. YOLO-UniOW incorporates Adaptive Decision Learning to replace computationally expensive cross-modality fusion with lightweight alignment in the CLIP latent space, achieving efficient detection without compromising generalization. Additionally, we design a Wildcard Learning strategy that detects out-of-distribution objects as “unknown” while enabling dynamic vocabulary expansion without the need for incremental learning. This design empowers YOLO-UniOW to seamlessly adapt to new categories in open-world environments. Extensive experiments validate the superiority of YOLO-UniOW, achieving 34.6 AP and 30.0 AP_r on LVIS with an inference speed of 69.6 FPS. The model also sets benchmarks on M-OWODB, S-OWODB, and nuScenes datasets, showcasing its unmatched performance in open-world object detection. Code and models are available at <https://github.com/THU-MIG/YOLO-UniOW>.

1. Introduction

Object detection has long been one of the most fundamental and widely applied techniques in the field of computer vision, with extensive applications in security [46], autonomous driving [57], and medical imaging [13]. Many remarkable works have achieved breakthroughs for object

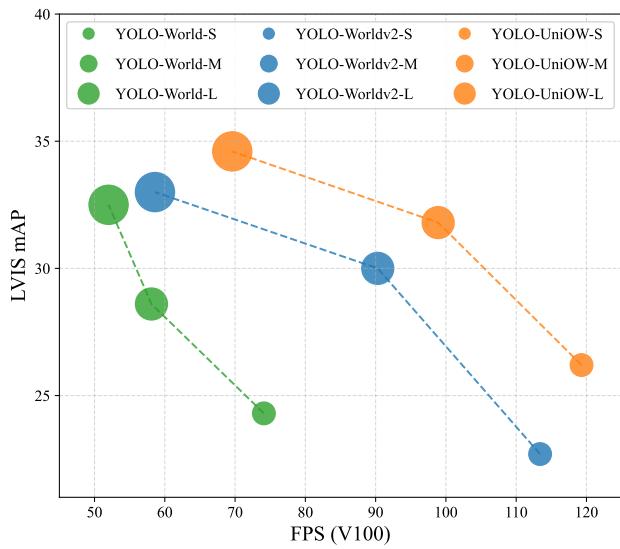


Figure 1. **Speed-Accuracy Trade-off Curve.** Comparison of YOLO-UniOW and recent methods in speed and accuracy on the LVIS minival dataset. Inference speed is measured on a single NVIDIA V100 GPU without TensorRT. Circle size indicates model size.

detection, such as Faster R-CNN [41], SSD [30], RetinaNet [26], etc.

In recent years, the YOLO (You Only Look Once) [1, 20, 40, 51] series of models has gained widespread attention for its outstanding detection performance and real-time efficiency. The recent YOLOv10 [51] establishes a new standard for object detection by employing a consistent dual assignment strategy, achieving efficient NMS-free training and inference.

However, traditional YOLO-based object detection models are often confined to a closed set definition, where objects of interest belong to a predefined set of categories.

In practical open-world scenarios, when encountering *unknown* categories that have not been encountered in the training datasets, these objects are often misclassified as background. This inability of models to recognize novel objects can also negatively impact the accuracy of known

*Equal contribution. ✉Corresponding author.

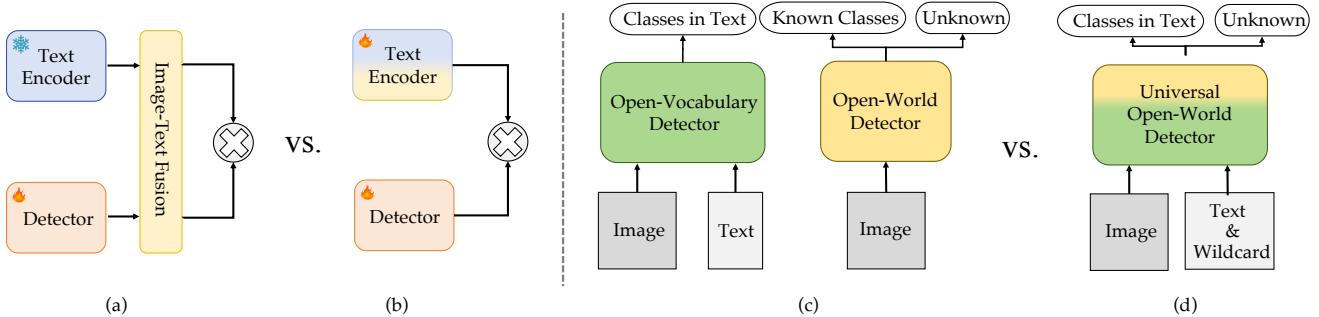


Figure 2. **Comparisons of Detection Framework.** (a) Open-vocabulary detector with cross-modality fusion. (b) Our efficient open-vocabulary detector with Adaptive Decision Learning. (c) Open-world and open-vocabulary detectors. (d) Our Uni-OWD detector for both open-vocabulary and open-world tasks.

categories, limiting their robust application in real-world scenarios.

Thanks to the development of vision-language models, such as [3, 19, 39, 47], combining their open-vocabulary capabilities with the efficient object detection of YOLO presents an appealing and promising approach for real-time open-world object detection. YOLO-World [4] is a pioneering attempt, where YOLOv8 [20] is used as the object detector, and CLIP’s text encoder is integrated as an open-vocabulary classifier for region proposals (i.e., anchors in YOLOv8). The decision boundary for object recognition is derived from representations of class names generated by CLIP’s text encoder. Additionally, a vision-language path aggregation network (RepVL-PAN) using re-parameterization [6, 50] is introduced to comprehensively aggregate text and image features for better cross-modality fusion.

Although YOLO-World is effective for open-vocabulary object detection (OVD), it still relies on a predefined vocabulary of class names, which must include all categories that are expected to be detected. This reliance significantly limits its ability to dynamically adapt to newly emerging categories, as determining unseen class names in advance is inherently challenging, preventing it being truly open-world. Moreover, the inclusion of RepVL-PAN introduces additional computational costs, especially with large vocabulary sizes, making it less efficient for real-world applications.

In this work, we first advocate a new setting of **Universal Open-World Object Detection (Uni-OWD)**, in which we encourage realizing open-world object detection (OWOD) and open-vocabulary object detection (OVD) with one unified model. Specifically, it emphasizes that the model can not only recognize categories unseen during training but also effectively classify unknown objects as “unknown”. Additionally, we call for an efficient solution following YOLO-World to meet the efficiency requirement in real-world applications. To achieve these, we propose a **YOLO-**

UniOW model to achieve effective universal open-world detection but also enjoying greater efficiency.

Our YOLO-UniOW emphasizes several insights for efficient Uni-OWD. (1) **Efficiency.** Except using recent YOLOv10 [51] for the more efficient object detector, we introduce a novel adaptive decision learning strategy, dubbed AdaDL, to wipe out the expensive cross-modality vision-language aggregation in RepVL-PAN, as illustrated in Fig. 2 (b). The goal of AdaDL is to adaptively capture task-related decision representations for object detection without sacrificing the generalization ability of CLIP. Therefore, we can well align the image feature and class feature directly in the latent CLIP space with no any heavy cross-modality fusion operations, achieving efficient and outstanding detection performance (see Fig. 1). (2) **Versatility.** The challenge of open-world object detection (OWOD) lies in differentiating all unseen objects with only one “unknown” category *without any supervision about unknown objects*. To solve this issue, we design a wildcard learning method that uses a wildcard embedding to unlock generic power of open-vocabulary model. This wildcard embedding is optimized through a simple self-supervised learning, which seamlessly adapts to dynamic real-world scenarios. As shown in Fig. 2 (d), our YOLO-UniOW can not only benefit from the dynamic expansion of the known category set like YOLO-World, *i.e.*, open-vocabulary detection, but also can highlight any out-of-distribution objects with “unknown” category for open-world detection. (3) **High performance.** We evaluated our zero-shot open-vocabulary capability in LVIS [14], and the open-world approach in benchmarks such as M-OWODB [44], S-OWODB [16], and nuScenes [2]. Experimental results show that our method can significantly outperform existing state-of-the-art methods for efficient OVD, achieving 34.6 AP, 30.0 AP_r, on the LVIS dataset with the speed of 69.6 FPS. Besides, YOLO-UniOW can also perform well in both zero-shot and task-incremental learning for open-world evaluation. These well demonstrate the effectiveness of the proposed YOLO-

UniOW.

The contributions of this work are as follows:

- We advocate a new setting of Universal Open-World Object Detection, dubbed Uni-OWD to solve the challenges of dynamic object categories and unknown target recognition with one unified model. We provide an efficient solution based on YOLO detector, ending up with our YOLO-UniOW.
- We design a novel adaptive decision learning (AdaDL) strategy to adapt the representation of decision boundary into the task of Uni-OWD without sacrificing the generalization ability of CLIP. Thanks to AdaDL, we can leave out the heavy computation of cross-modality fusion operation used in previous works.
- We introduce wildcard learning to detect unknown objects, enabling iterative vocabulary expansion and seamless adaptation to dynamic real-world scenarios. This strategy eliminates the reliance on incremental learning strategies.
- Extensive experiment across benchmark for both open-vocabulary object detection and open-world object detection show that YOLO-UniOW can significantly outperform existing methods, well demonstrating its versatility and superiority.

2. Related Work

2.1. Open-Vocabulary Object Detection

Open-Vocabulary Object Detection (OVD) has emerged as a prominent research direction in computer vision in recent years. Unlike traditional object detection, OVD enables the detection dynamically expand categories without relying heavily on the fixed set of categories defined in the training dataset. Several works have explored leveraging Vision-Language Models (VLMs) for enhancing object detection. For instance, [4, 24, 28, 32, 42, 59, 60, 65, 68] utilize large-scale, easily accessible text-image pairs for pre-training, resulting in more robust and generalizable detectors, which are subsequently fine-tuned on specific target datasets. In parallel, [12, 36, 38, 53] focus on distilling the alignment of visual-text knowledge from VLMs into object detection, emphasizing the design of distillation losses and the generation of object proposals. Additionally, [7, 11, 54] investigate various prompt modeling techniques to more effectively transfer VLM knowledge to the detector, enhancing its performance in open-vocabulary and unseen category tasks.

2.2. Open-World Object Detection

Open-World Object Detection (OWOD) is an emerging direction in object detection, aiming to address the challenge of dynamic category detection. The goal is to enable detection models to identify known categories while recognizing

unknown categories, and to incrementally adapt to new categories over time. Through methods such as manual annotation or active learning [31, 43, 62], unknown categories can be progressively converted into known categories, facilitating continuous learning and adaptation.

The concept of OWOD was first introduced by Joseph et al. [21], whose framework relies on incremental learning. By incorporating an energy-based object recognizer into the detection head, the model gains the ability to identify unknown categories. However, this method depends on replay mechanisms, requiring access to historical task data to update the model. Additionally, it often exhibits a bias toward known categories when handling unknown objects, limiting its generalization capabilities. To address these limitations, many subsequent studies have been proposed. For instance, [35, 67] improved the experimental setup for OWOD by introducing more comprehensive benchmark datasets and stricter evaluation metrics, enhancing the robustness of unknown category detection. While these improvements achieved promising results in controlled experimental settings, their adaptability to complex scenarios and dynamic category changes remains inadequate. Recent research has shifted focus toward optimizing the feature space to better separate known and unknown categories. Methods such as [9, 48, 55, 61] propose advancements in feature space extraction, enabling models to more effectively extract feature information for the localization and identification of unknown objects. Recently, several methods [25, 34, 71] have emerged, leveraging pretrained models for open-world object detection and achieving significant improvements.

2.3. Parameter Efficient Learning

Prompt learning has emerged as a significant research direction in both natural language processing (NLP) and computer vision. By providing carefully designed prompts to pre-trained large models such as [39], prompt learning enables models to perform specific tasks in unsupervised or semi-supervised settings efficiently. Methods such as [17, 23, 56, 58, 69, 70] introduce learnable prompt embeddings, moving beyond fixed, handcrafted prompts to enhance flexibility across various visual downstream tasks. And DetPro [7] is the first to apply it to open-vocabulary object detection, achieving significant improvements using learnable prompts derived from text inputs.

Low-Rank Adaptation (LoRA) [18] and its derivatives [29, 63, 64], as a parameter-efficient fine-tuning technique, has demonstrated outstanding performance in adapting large models. By inserting trainable low-rank decomposition modules into the weight matrices of pre-trained models without altering the original weights, LoRA significantly reduces the number of trainable parameters. CLIP-LoRA [63] introduces LoRA into VLM models as a re-

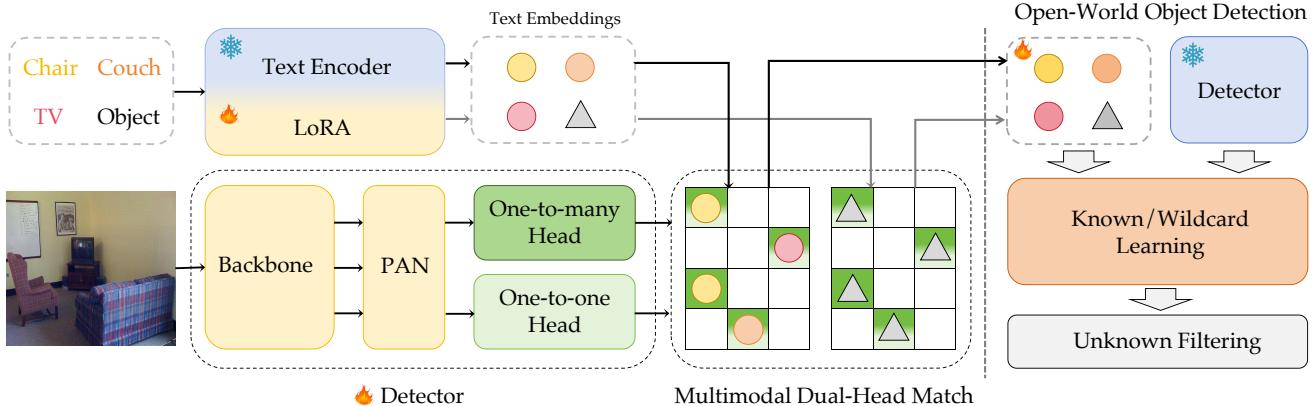


Figure 3. Our Proposed Efficient Universal Open-World Object Detection Pipeline. Open-Vocabulary Pretraining (left): Using a *Multimodal Dual-Head Match* for efficient end-to-end object detection, *AdaDL* in text encoder for adaptive decision boundary learning. Open-World Fine-tuning (right): Utilizing calibrated text embeddings and the detector to adaptively detect both known and unknown objects with the assistance of the wildcard. A filtering strategy is employed to remove duplicate unknown predictions, ensuring efficient and effective open-world object detection.

placement for adapters and prompts, enabling fine-tuning for downstream tasks with faster training speeds and improved performance.

3. Efficient Universal Open-World Object Detection

3.1. Problem Definition

Universal Open-World Object Detection (Uni-OWD) extends the challenges of Open Vocabulary Detection (OVD) and Open-World Object Detection (OWOD), aiming to create a unified framework that not only detects known objects in the vocabulary but also dynamically adapts to unknown objects while maintaining scalability and efficiency in real-world scenarios.

Define the object category set as $\mathcal{C} = \mathcal{C}_k \cup \mathcal{C}_{unk}$, where \mathcal{C}_k represents the set of known categories, \mathcal{C}_{unk} represents the set of unknown categories, and $\mathcal{C}_k \cap \mathcal{C}_{unk} = \emptyset$. Given an input image \mathcal{I} and a vocabulary \mathcal{V} , the goal of Uni-OWD is to design a detector \mathcal{D} that satisfies the following objectives:

1. For each category $c_k \in \mathcal{C}_k$, represented by its text $\mathcal{T}_{c_k} \in \mathcal{V}$, the detector \mathcal{D} should accurately predict the bounding boxes \mathcal{B}_{c_k} and their associated category labels c_k by $\mathcal{D}(\mathcal{I}, \mathcal{V}) \rightarrow \{(b, c_k) \mid b \in \mathcal{B}_{c_k}, c_k \in \mathcal{C}_k\}$
2. For objects belonging to \mathcal{C}_{unk} , the detector should identify their bounding boxes \mathcal{B}_{unk} and assign them the generic label ‘‘unknown’’ with a wildcard \mathcal{T}_w , such that: $\mathcal{D}(\mathcal{I}, \mathcal{T}_w) \rightarrow \{(b, \text{unknown}) \mid b \in \mathcal{B}_{unk}\}$
3. The detector can iteratively expand the known category set \mathcal{C}_k and vocabulary \mathcal{V} by discovering new categories \mathcal{C}_{new} from \mathcal{C}_{unk} , represented as $\mathcal{C}_k^{t+1} = \mathcal{C}_k^t \cup \mathcal{C}_{new}$

The Uni-OWD framework is designed to develop a detector that leverages a textual vocabulary and a wildcard to

identify both known and unknown object categories within an image, combining the strengths of open-vocabulary and open-world detection tasks. It ensures precise detection and classification for known categories while assigning a generic ‘‘unknown’’ label to unidentified objects. This design promotes adaptability and scalability, making it well-suited for dynamic and real-world applications.

3.2. Efficient Adaptive Decision Learning

Designing a universal open-world object detection model suitable for deployment on edge and mobile devices demands a strong emphasis on efficiency. Traditional open-vocabulary detection models [4, 28, 42, 65] align text and image modalities by introducing fine-grained fusion operations in the early layers. Then they rely on contrastive learning for both modalities to establish decision boundaries for object classification, enabling the model to adapt dynamically to novel classes during inference by leveraging new textual inputs.

YOLO-World [4] proposed an efficient architecture, RepVL-PAN, to perform image-text fusion through re-parameterization. Despite its advancements, the model’s inference speed is still heavily influenced by the number of textual class inputs. This poses a challenge for low-compute devices, where performance degrades sharply as the number of text inputs increases, making it unsuitable for real-time detection tasks in complex, multi-class scenarios. To address this, we propose a adaptive decision learning strategy (*AdaDL*) to eliminate the heavy early-layer fusion operation.

During the construction of decision boundaries, most existing methods freeze the text encoder and rely on pre-trained models, such as BERT[5] or CLIP[39], to extract

textual features for interaction with visual features. Without a fusion structure, the text features struggle to capture image-related information dynamically, leading to suboptimal multimodal decision boundary construction when adjustments are made solely to the image features. To overcome this, our AdaDL strategy aims to enhance the decision representation during training for the Uni-OWD scenario. Specifically, during training, we introduce efficient parameters into the text encoder by incorporating Low-Rank Adaptation (LoRA) into all query, key, value and output projection layers, which can be described as:

$$h = W'x = W_0x + \Delta Wx \quad (1)$$

where W_0 represents the pretrained weights of the CLIP text encoder, and ΔW is the product of two low-rank matrices. The model’s input and output are x and h . The rank is set to a value much smaller than the model’s feature dimension. This strategy ensures that the pre-trained parameters of the text encoder remain unchanged while low-rank matrices dynamically store information related to cross-modality interactions during training. By continuously calibrating the outputs of text encoder, this method allows the decision boundaries constructed by both modalities to adapt more effectively to each other. In practice, the calibrated text embeddings can be precomputed and stored offline during inference, thereby avoiding the computational cost of the text encoder.

YOLOv10 as the efficient object detector. To improve the efficiency, we accommodate the proposed adaptive decision learning strategy into the recent advanced YOLOv10 [51] as the efficient object detector. We employ a multimodal dual-head match to adapt the decision boundary for both classification heads in YOLOv10. Specifically, during the region-text contrastive learning between the region anchor and the class text, we refine the region embeddings from two heads by aligning them with shared, semantically rich text representations, enabling seamless end-to-end training and inference. Furthermore, we integrate a consistent dual alignment strategy for region contrastive learning, where the dual-head matching process is formalized as:

$$m(\alpha, \beta) = s^\alpha \times u^\beta \quad (2)$$

where u represents the IoU value between the predicted box and the ground truth box. s is the classification score obtained by multi-modal information, which is derived as:

$$s = \text{sim}(I, T) \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, T is the embeddings from the text $\mathcal{T} \in \mathcal{V}$ and I is the pixel-level features from image \mathcal{I} . To ensure minimal supervision gap between the both heads during multimodel dual-head matching, we adopt the consistent settings, where $\alpha_{o2o} = \alpha_{o2m}$

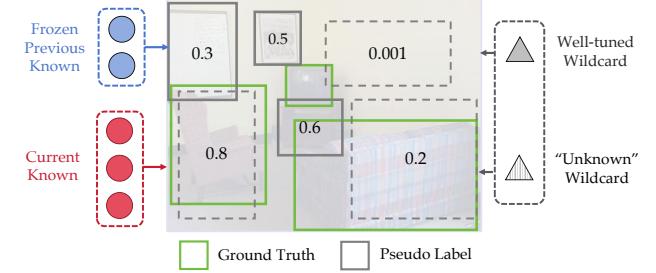


Figure 4. **The Process of Known/Wildcard Learning.** The text embeddings for previously known classes are frozen, while the embeddings for currently known classes are fine-tuned using ground truth labels. The “unknown” wildcard is supervised by pseudo labels generated by the well-tuned wildcard predictions. It shows well-tuned wildcard’s prediction scores and the boxes with low confidence scores or high IoU values (dashed boxes) with known class ground truth are filtered out.

and $\beta_{o2o} = \beta_{o2m}$. This allows the one-to-one head to effectively learn consistent supervisory signals with one-to-many head.

As a result, the calibrated text encoder and YOLO structure can operate entirely independently in the early stages, eliminating the need for fusion operations while efficiently adapting to better multimodal decision boundaries.

3.3. Open-World Wildcard Learning

In the previous section, we introduced the *AdaDL* to improve the efficiency of open-vocabulary object detection, mitigating the impact of large input class text on inference latency meanwhile improves its performance. This strategy enables real-world applications to expand the vocabulary while maintaining high efficiency, covering as many objects as possible. However, open-vocabulary models inherently rely on predefined vocabularies to detect and classify objects, which limits their capability in real-world scenarios. Some objects are difficult to predict or describe using textual inputs, making it challenging for open-vocabulary models to detect these out-of-vocabulary instances.

To address this, we propose a *wildcard learning* approach that enables the model to detect objects not present in the vocabulary and label them as “unknown” rather than ignoring them. Specifically, we directly leverage a wildcard embedding to unlock generic power of open-vocabulary model. As shown in Tab. 4, after the decision adaptation, the wildcard \mathcal{T}_w (e.g. “object”) demonstrates remarkable capability in capturing unknown objects within a scene in a zero-shot manner. To further enhance its effectiveness, we fine-tune its text embedding on the pretraining dataset for a few epochs. During this process, all ground-truth instances are treated as belonging to the same “object” class. This fine-tuning enables the embedding to capture richer semantic

tics, empowering the model to identify objects that might have been overlooked by the predefined specific classes.

To avoid duplicate predictions for known classes, we utilize this well-tuned wildcard embedding T_{obj} to teach an “unknown” wildcard embedding T_{unk} . The “unknown” wildcard is trained in a self-supervised manner without relying on ground-truth labels of “unknown” class. As shown in Fig. 4, predictions that has the highest similarity score with T_{obj} across all known classes embeddings are used as pseudo label candidates. To further refine these candidates, we introduce a simple selection process:

$$\Phi(s, u) = \begin{cases} 1, & \text{if } (u < \sigma_1) \wedge (s > \sigma_2) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where u is the maximum IoU between predictions and known class ground truth boxes. And the predictions with u below a threshold σ_1 or classification score s above a threshold σ_2 are selected. The remaining predictions are assigned to T_{unk} as target labels.

For known classes, only their corresponding text embeddings T_k from $\mathcal{T}_k \in \mathcal{V}$ are fine-tuned on downstream tasks by multimodal dual-head match to enhance their similarity scores s_k aligned with target score o_k . These embeddings are subsequently frozen to preserve performance and avoid degradation when new classes are introduced. Unlike traditional open-world methods [15, 21] relying on exemplar replay to do incremental learning, our method can avoid catastrophic forgetting without extra exemplars, as each text embeddings are fine-tuned independently.

Since T_k , T_{obj} and T_{unk} calculate similarity scores only in frozen classification head, there is no loss from box regression, focusing exclusively on learning class-specific information. The soft target scores of T_{unk} are directly derived from the similarity scores s_{obj} from the T_{obj} . Therefore, the fine-tuning loss is formulated as the combination of the current known loss and the unknown loss, ensuring the model learns effectively from both known and unknown categories during training:

$$\mathcal{L} = \mathcal{L}_k(s_k, o_k) + \Phi(s_{obj}, u_{obj}) \cdot \mathcal{L}_{unk}(s_{unk}, s_{obj}) \quad (5)$$

where s_{unk} is the prediction score from “unknown” wildcard. \mathcal{L} represents the binary cross-entropy (BCE) loss. During inference, we employ a simple and efficient unknown filtering strategy \mathcal{F} for unknown class predictions P_{unk} that have a high IoU with confident known class predictions P_k to further de-duplicate.

$$\mathcal{F}(P_{unk}) = \{p_u \in P_{unk} \mid \text{IoU}(p_u, p_k) < \tau, \forall p_k \in P_k\} \quad (6)$$

where τ is the IoU threshold for unknown filtering.

Subsequently, new categories can be discovered from the predictions of the unknown class, and their class names will be added to the vocabulary \mathcal{V} , where they serve as known classes for the next iteration.

4. Experiments

4.1. Dataset

We evaluate our method on two distinct setups, targeting both OVD and OWOD. Our experiments leverage diverse datasets to comprehensively assess the model’s performance in detecting known and unknown objects.

Open-Vocabulary Object Detection: For open vocabulary detection, the model is trained on a combination of Objects365 [45] and GoldG [22] datasets, and evaluated on the LVIS [14] dataset. The LVIS dataset contains 1,203 categories, exhibiting a realistic long-tailed distribution of rare, common, and frequent classes. This setup focuses on evaluating the model’s capacity to align visual and language representations, detect novel and unseen categories, and generalize across a large-scale, long-tailed dataset.

Open-World Object Detection: For open-world object detection, we evaluate our method on three established OWOD benchmarks: **M-OWODB**: This benchmark combines the COCO [27] and PASCAL VOC [8] datasets, where known and unknown classes are mixed across tasks. It is divided into four sequential tasks. At each task, the model learns new classes while the remaining classes remain unknown. **S-OWODB**: Based solely on COCO, this benchmark separates known and unknown classes by their superclass. **nu-OWODB**: This benchmark is derived from [25] and based on the nuSences dataset [2]. This benchmark is specifically designed to evaluate the model’s capability in autonomous driving scenarios. The nu-OWODB captures the complexity of urban driving environments, including crowded city streets, challenging weather conditions, frequent occlusions, and dense traffic with intricate interactions between objects.

By incorporating these benchmarks, we assess the model’s ability to handle real-world OWOD challenges while maintaining robustness and scalability across diverse settings.

4.2. Evaluation Metrics

Open-Vocabulary Evaluation: Similar to YOLO-World and other pre-trained models, we evaluate the zero-shot capability of our pre-trained model on the LVIS minival dataset, which contains the same images in COCO validation set. For fair and consistent comparison, we use *standard AP* metrics to measure the model’s performance.

Open-World Evaluation: We adapt the pretrained open-vocabulary model to the open-world scenario, enabling it to recognize both known and unknown objects. For known objects, we use *mAP* as the evaluation metric. To further assess catastrophic forgetting during incremental tasks, the *mAP* is divided into previous known (PK) and current known (CK) categories. For unknown objects, since it is impractical to exhaustively annotate all remaining ob-

Table 1. **Zero-shot evaluation on the LVIS minival dataset.** The AP of our model is presented for both the one-to-one head (left) and the one-to-many head (right). All speed measurements are conducted on a V100 GPU using PyTorch without TensorRT. FPS^f denotes the speed in the forward process without post-processing. We limit number of prediction boxes under 3000 before NMS for FPS measurement.

Model	Backbone	Pre-trained Data	Params	AP	AP_r	AP_c	AP_f	FPS	FPS^f
GLIPv2-T	Swin-T	O365,GoldG,Cap4M	232M	29.0	-	-	-	0.12	-
Grounding DINO 1.5 Edge	EfficientViT-L1	Grounding-20M	-	33.5	28.0	34.3	33.9	-	-
OmDet-Turbo-T	Swin-T	O365,GoldG	-	30.3	-	-	-	-	-
YOLO-World-S	YOLOv8-S	O365,GoldG	13M	24.3	16.6	22.1	27.7	-	74.1
YOLO-World-M	YOLOv8-M	O365,GoldG	29M	28.6	19.7	26.6	31.9	-	58.1
YOLO-World-L	YOLOv8-L	O365,GoldG	48M	32.5	22.3	30.6	36.1	-	52.0
YOLO-Worldv2-S	YOLOv8-S	O365,GoldG	13M	22.7	16.3	20.8	25.5	87.3	113.4
YOLO-Worldv2-M	YOLOv8-M	O365,GoldG	29M	30.0	25.0	27.2	33.4	74.6	90.3
YOLO-Worldv2-L	YOLOv8-L	O365,GoldG	48M	33.0	22.6	32.0	35.8	51.2	58.6
YOLO-UniOW-S	YOLOv10-S	O365,GoldG	7.5M	26.2/27.4	24.1/26.0	24.9/25.6	27.7/29.3	98.3	119.3
YOLO-UniOW-M	YOLOv10-M	O365,GoldG	16.2M	31.8/32.8	26/26.6	30.5/31.8	34/34.9	86.2	98.9
YOLO-UniOW-L	YOLOv10-L	O365,GoldG	29.4M	34.6/34.8	30/34.2	33.6/32.4	36.3/37.0	64.8	69.6

jects in the scene, we employ the *Recall* metric to evaluate the model’s ability to detect unknown categories. Additionally, WI [21] and A-OSE [21] are used to measure the extent to which unknown objects interfere with known object predictions. However, due to their instability, these metrics are provided for reference purposes only.

4.3. Implementation Details

Open-Vocabulary Detection: Our image detector follows YOLOv10 [51], which provides an efficient design for dual-head training. Similar to YOLO-World [4], we utilize a pre-trained CLIP text encoder. However, we do not perform image-text fusion in the neck. Instead, we align the two modalities solely in the head using efficient adaptive decision learning. During pretraining, we incorporate low-rank matrices into the all projection layers in the CLIP text encoder. The rank for the matrices are set to 16. Our pre-training is conducted on 8 GPUs, with a batch size of 128. Both the YOLO model and LoRA parameters for the text encoder are trained with an initial learning rate of 5×10^{-4} and weight decay of 0.025.

Open-World Detection: All the wildcard embeddings are initialized from text features extracted from a generic text, “object”, by our calibrated text encoder. We use the same training datasets employed for open-vocabulary pre-training to fine-tune the wildcard embedding T_{obj} . Specifically, the wildcard embedding is trained for 3 epochs with a learning rate of 1×10^{-4} . Using the well-tuned wildcard as an anchor, the learning rate for fine-tuning the known and unknown class embeddings is set to 1×10^{-3} , with weight decay set to 0. All the other parts of model are frozen and mosaic augmentation is not applied during this stage.

For training the “unknown” wildcard, pseudo-labels are

selected based on an IoU threshold $\sigma_1 = 0.5$ and a score threshold $\sigma_2 = 0.01$. During inference, known class predictions with score greater than 0.2 are confident predictions, and $\tau = 0.99$. For known class detection, predictions with scores below 0.05 are filtered out as default.

All fine-tuning experiments are conducted on 8 GPUs, with a batch size of 16 per GPU. Notably, all open-world experiments are evaluated using the one-to-one head, which not requires NMS operations for post-processing.

4.4. Quantitative Results

Tab. 1 demonstrates that model with efficient adaptive decision learning achieves significant zero-shot performance improvements on the LVIS benchmark, outperforming recent real-time state-of-the-art open-vocabulary models [4, 42, 66]. For the small model (-S), we observe that using predictions from the one-to-one head alone improves the detection performance for rare classes by 6.4% and common classes by 3.2%. Furthermore, employing a one-to-many head structure with NMS achieves even greater performance gains. This clearly demonstrates that in previous pretraining processes, the multimodal decision boundaries were fully constructed by incorporating AdaDL. Additionally, leveraging the efficient model architecture and the nature of end-to-end detection, our approach gains faster speed and eliminates the need for NMS during inference, making it highly efficient for real-world applications.

To address open-world demands, we adapt our well-adapted open-vocabulary model to recognize unknown classes that are not present in the predefined vocabulary through wildcard learning. As shown in Tab. 2, the open-vocabulary model demonstrates outstanding performance in open-world scenarios due to its rich knowledge. Through

Table 2. **OWOD results on M-OWODB (top) and S-OWODB (bottom).** Comparision of unknown class recall (U-Recall) and mean average precision (mAP) for known classes. Our method outperforms both traditional models and those leveraging pretrained knowledge. OVOW* represents the reproduced version using YOLO-Worldv2-S to ensure a fair comparison with our model at the same scale.

Task IDs (\rightarrow)	Task1			Task 2			Task3			Task 4			
	Methods	U-Recall (\uparrow)	mAP (\uparrow) CK	U-Recall (\uparrow)			mAP (\uparrow)			U-Recall (\uparrow)	mAP (\uparrow)		
				PK	CK	Both	PK	CK	Both		PK	CK	Both
M-OWODB													
ORE-EBUI [21]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
OW-DETR [15]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
PROB [72]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
CAT [33]	23.7	60.0	19.1	55.5	32.2	44.1	24.4	42.8	18.8	34.8	34.4	16.6	29.9
RandBox [52]	10.6	61.8	6.3	-	-	45.3	7.8	-	-	39.4	-	-	35.4
EO-OWOD [49]	24.6	61.3	26.3	55.5	38.5	47	29.1	46.7	30.6	41.3	42.4	24.3	37.9
MEPU-FS [10]	31.6	60.2	30.9	57.3	33.3	44.8	30.1	42.6	21.0	35.4	34.8	18.9	30.4
MAVL [37]	50.1	64.0	49.5	61.6	30.8	46.2	50.9	43.8	22.7	36.8	36.2	20.6	32.3
SKDF [34]	39.0	56.8	36.7	52.3	28.3	40.3	36.1	36.9	16.4	30.1	31.0	14.7	26.9
OVOW* [25]	65.9	57.1	72.7	55.1	38.6	47.0	66.48	43.4	24.3	37.0	36.2	20.4	32.30
YOLO-UniOW-S	80.6	70.4	80.8	70.4	42.9	56.6	79.9	56.6	33.1	48.8	48.8	26.7	43.3
YOLO-UniOW-M	82.6	73.6	82.6	73.4	48.4	60.9	81.5	60.9	39.0	53.6	53.6	32.0	48.2
S-OWODB													
ORE-EBUI [21]	1.5	61.4	3.9	56.7	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
OW-DETR [15]	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
PROB [72]	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
CAT [33]	24.0	74.2	23.0	67.6	35.5	50.7	24.6	51.2	32.6	45.0	45.4	35.1	42.8
EO-OWOD [49]	24.6	71.6	27.9	64	39.9	51.3	31.9	52.1	42.2	48.8	48.7	38.8	46.2
MEPU-FS [10]	37.9	74.3	35.8	68.0	41.9	54.3	35.7	50.2	38.3	46.2	43.7	33.7	41.2
SKDF [34]	60.9	69.4	60.0	63.8	26.9	44.4	58.6	46.2	28.0	40.1	41.8	29.6	38.7
OVOW* [25]	69.8	52.7	73.2	54.1	37.1	45.1	69.7	42.3	31.6	38.7	38.6	29.5	36.4
YOLO-UniOW-S	82.2	69.20	81.4	69.2	50.64	59.4	81.0	59.4	44.4	54.4	54.4	44.7	52.0
YOLO-UniOW-M	84.5	74.4	83.4	74.4	56.9	65.2	83.0	65.2	52.2	61.0	61.0	52.7	58.9

Table 3. **Evaluation on nu-OWODB.** Our method outperforms all other approaches in unknown metrics, demonstrating its strong adaptability for real-world applications. OVOW* represents the reproduced model based on YOLO-Worldv2-S

Task IDs (\rightarrow)	Task 1			Task 2			Task3								
	Methods	WI (\downarrow)	A-OSE (\downarrow)	U-Recall (\uparrow)	mAP (\uparrow) CK	WI (\downarrow)	A-OSE (\downarrow)	U-Recall (\uparrow)	mAP (\uparrow)	PK	CK	Both	mAP (\uparrow)		
										PK	CK	Both			
PROB [72]													18.1	16	17.5
EO-OWOD [49]	0.0059	223	1.4	22.4	25.1	0.0015	1583	2.8	27.2	6.7	18.8	21.8	25.6	22.8	
OVOW* [25]	0.0080	16478	16.7	14.2	22.4	0.0096	6394	21.74	13.6	6.0	10.5	10.0	9.7	9.9	
YOLO-UniOW-S	0.0137	1658	37.5	21.5	0.0074	1265	30.0	21.5	9.8	16.7	16.7	15.6	16.4		
YOLO-UniOW-M	0.0147	1722	41.8	24.8	0.0067	1156	35.4	24.8	13.1	20.0	20.0	18.4	19.6		

our wildcard learning strategy, the model achieves superior performance in both unknown and known class recognition compared to traditional open-world methods. Moreover, it outperforms recent open-world detection models that leverage pre-training models [10, 25, 34]. Notably, our simpler and more efficient approach surpasses the state-of-the-art OVOW model [25], which is also based on YOLO-World structure. Our method achieves a significant improvement

in unknown recall and known mAP, demonstrating its effectiveness and robustness in open-world detection tasks. Furthermore, we evaluated the model’s capability in real-world autonomous driving scenarios. As shown in Tab. 3, our model, using a simpler approach, achieves superior unknown detection performance compared to the other methods.

Benefiting from AdaDL and wildcard learning strategies,

Table 4. Comparison with Oracle and Zero-shot Settings. We compare our method with the closed-set YOLOv10 model trained in an oracle manner, and our pre-trained open-vocabulary model in a zero-shot setting. Our method achieves great improvement, even surpassing the close-set oracle model.

Task IDs (\rightarrow)	Task 1						Task 2						Task 3															
	Methods			WI (\downarrow)	A-OSE (\downarrow)	U-Recall (\uparrow)	mAP (\uparrow)	CK	Methods			WI (\downarrow)	A-OSE (\downarrow)	U-Recall (\uparrow)	mAP (\uparrow)	PK	CK	Both	Methods			WI (\downarrow)	A-OSE (\downarrow)	U-Recall (\uparrow)	mAP (\uparrow)	PK	CK	Both
YOLOv10-S-oracle	0.0482	18447	61.5	64.6	0.0147	12614	84.2	64.3	48.3	56.3	0.0111	10795	83.9	55.9	31.1	47.6												
YOLO-UniOW-S-zs	0.0287	2827	48.8	68.1	0.0185	2173	50.5	68.1	40.1	54.1	0.0125	1775	52.1	53.6	30.7	46.0												
YOLO-UniOW-M-zs	0.0232	2409	54.4	71.0	0.0130	1740	57.6	70.8	44.2	57.5	0.00888	1448	59.3	57.7	37.0	50.8												
YOLO-UniOW-S	0.0229	1609	80.6	70.4	0.0133	1208	80.8	70.4	42.9	56.6	0.0091	1049	79.9	56.6	33.1	48.8												
YOLO-UniOW-M	0.0210	1514	82.6	73.4	0.01093	1027	82.62	73.4	48.4	60.9	0.00723	872	81.5	60.9	39.0	53.6												

our model captures a broader range of unknown objects through wildcard embeddings while maintaining accurate recognition of known categories. Notably, as the model scales up, the capability of model to detect known and unknown objects increases progressively, which shows the effectiveness of our methods at different model scale.

4.5. Ablation Study

Open-Vocabulary Detection: We conducted a series of ablation studies on the small scale model to evaluate the impact of image-text fusion. Due to differences in experimental settings, we first reproduced YOLO-Worldv2-S under our setup. Interestingly, as shown in Tab. 5 our findings reveal that a smaller batch size and learning rate yield better pretraining performance, particularly improving detection for frequent classes by 2.2%. Building on this, we removed the VL-PAN structure and observed that the model’s detection accuracy remains largely unaffected. Notably, it demonstrated improved generalization for rare classes. Replacing YOLO-World’s YOLOv8 structure with YOLOv10 and using a dual-head match demonstrated that the one-to-many head benefits more from these changes, achieving improved performance over YOLO-World. However, the one-to-one head still struggled with alignment, particularly in rare class detection. To address this, we calibrate text encoder with AdaDL, making both image and text encoder learn decision boundaries simultaneously, which attains significant improvements.

As shown in Tab. 6, we compare the different methods for AdaDL to calibrate text encoder. Performing full fine-tuning improves overall accuracy but reduces performance on rare classes, likely due to overfitting. We assume this is caused by the large gap between the number of image and text training parameters. Introducing parameter-efficient methods like prompt tuning [70] and deep prompt tuning [23] significantly improved alignment, enabling the one-to-one head to match the one-to-many head in performance. And as the training parameters increase, the performance also improves. Finally, using LoRA for text encoder across all projection layers further adapt text information to be region-aware. This approach yielded the best overall

Table 5. Ablations on Pre-training Settings. * means reproduced version in our experiment setting. w/o VL-PAN denotes eliminating RepVL-PAN structure in YOLO-Worldv2.

Method	AP	AP _r	AP _c	AP _f
YOLO-World	24.3	16.6	22.1	27.7
YOLO-Worldv2	22.7	16.3	20.8	25.5
YOLO-Worldv2*	23.5	16.7	20.2	27.7
w/o VL-PAN	22.8	17.1	19.4	26.8
+Dual-Head Match	21.5/23.3	12.4/17.3	18.7/20.4	25.6/26.9
+AdaDL	26.2/27.4	24.1/26.0	24.9/25.6	27.7/29.3

Table 6. Ablations on AdaDL Methods. We ablate different methods for adaptive decision learning.

Method	AP	AP _r	AP _c	AP _f
Frozen	21.5/23.3	12.4/17.3	18.7/20.4	25.6/26.9
Full Fine-tune	23.2/23.8	13.7/13.3	20.5/20.8	27.2/28.4
Prompt	24.1/25.1	18.5/19.1	21.9/23.2	27.2/28.0
Deep Prompt	24.5/25.4	19.5/18.5	22/23.4	27.6/28.5
LoRA	26.2/27.4	24.1/26.0	24.9/25.6	27.7/29.3

results and was adopted for our final experiments.

Open-World Detection: We compared the performance of close-set YOLOv10 trained with unknown class labels (oracle) and zero-shot performance of our open-vocabulary model on the M-OWODB dataset. The results show in Tab. 4, even in a zero-shot setting, our open-vocabulary model achieves higher known class accuracy than the oracle-trained YOLOv10 model. Moreover, when we only simply use vanilla “object” as text input, it achieves better Unknown recall than traditional owod methods, which further validates the effectiveness of our open-vocabualry methods. By applying our wildcard embeddings, the model’s unknown detection capability is fully unlocked, surpassing the performance of models trained with oracle supervision on unknown labels across different tasks. And as the model scales up, its ability to detect known and unknown class increases simultaneously.

4.6. Qualitative Result

For the open-vocabulary model, we input the 1,023 category names from the LVIS dataset as prompt, comparing the



Figure 5. Visualization Results on Zero-shot Inference on LVIS. We present visualization results with YOLO-Worldv2 both in *small* scale, using LVIS 1023 class names as text prompts. The model pretrained with our strategy demonstrates exceptional capability in detecting objects within complex scenes and recognizing a broader range of novel classes.

zero-shot performance on LVIS with YOLO-Worldv2, as shown in Fig. 5. It shows that our AdaDL strategy enhances the model’s decision boundaries to detect objects of varying sizes, distances, or those partially occluded, with higher confidence scores. Moreover, the improved alignment between visual and calibrated semantic information enables the model to correctly classify detected objects, capturing more diverse categories.

In Fig. 6, we compare the performance of an open-vocabulary model using text embeddings for all 80 known classes in the M-OWODB dataset with our model, which uses text embeddings for only half of the known classes (similar to the Task 2 scenario) and an extra “unknown” wildcard to detect unknown objects. The results demonstrate that our model not only identifies the remaining 40 unknown classes without corresponding text inputs but also detects additional objects. This indicates that the “unknown” wildcard effectively retains the rich semantic knowledge from pretraining while learning downstream task-specific knowledge, showcasing strong generalization capabilities that align with real-world requirements.

5. Conclusion

In this work, we propose Universal Open-World Object Detection (Uni-OWD), a new paradigm to tackle the challenges of dynamic object categories and unknown target recognition within a unified framework. To address

this, we introduce YOLO-UniOW, an efficient solution based on the YOLO detector. Our framework incorporates several innovative strategies: the Adaptive Decision Learning (AdaDL) strategy, which seamlessly adapts decision boundaries for Uni-OWD tasks, and Wildcard Learning, using a “unknown” wildcard embedding to enable the detection of unknown objects, supporting iterative vocabulary expansion without incremental learning. Extensive experiments across benchmarks for both open-vocabulary and open-world object detection validate the effectiveness of our approach. The results demonstrate that YOLO-UniOW significantly outperforms state-of-the-art methods, offering a versatile and superior solution for open-world object detection. This work highlights the potential of our framework for real-world applications, paving the way for further advancements in this evolving field.

References

- [1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020. 1
 - [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi-anicarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving, 2020. 2, 6
 - [3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval.



Figure 6. **Visualization Results on M-OWODB.** Compared to the open-vocabulary model using prompts with all 80 classes, our approach that extends to open-world only employs 40 class embeddings with an additional “unknown” wildcard.

- In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020. 2
- [4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2, 3, 4, 7
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 4
- [6] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 2
- [7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3
- [8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 6
- [9] Ruohuan Fang, Guansong Pang, Lei Zhou, Xiao Bai, and Jin Zheng. Unsupervised recognition of unknown objects for open-world object detection, 2023. 3
- [10] Ruohuan Fang, Guansong Pang, Lei Zhou, Xiao Bai, and Jin Zheng. Unsupervised recognition of unknown objects for open-world object detection. *arXiv preprint arXiv:2308.16527*, 2023. 8
- [11] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-detect: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 3
- [12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022. 3
- [13] Yuchen Guo, Yuwei He, Jinhao Lyu, Zhanping Zhou, Dong Yang, Liangdi Ma, Hao-tian Tan, Changjian Chen, Wei Zhang, Jianxing Hu, et al. Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study. *The Lancet Digital Health*, 4(8):e584–e593, 2022. 1
- [14] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 2, 6
- [15] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9235–9244, 2022. 6, 8
- [16] Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer, 2022. 2
- [17] Tianxiang Hao, Hui Chen, Yuchen Guo, and Guiguang Ding. Consolidator: Mergeable adapter with grouped connections for visual adaptation, 2023. 3
- [18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

- Lora: Low-rank adaptation of large language models, 2021. 3
- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [20] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics Yolov8. <https://github.com/ultralytics/ultralytics>, 2023. 1, 2
- [21] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection, 2021. 3, 6, 7, 8
- [22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding, 2021. 6
- [23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3, 9
- [24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3
- [25] Zizhao Li, Zhengkang Xiang, Joseph West, and Kourosh Khoshelham. From open vocabulary to open world: Teaching vision language models to detect novel objects, 2024. 3, 6, 8
- [26] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 1
- [27] 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. 6
- [28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 3, 4
- [29] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 3
- [30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. 1
- [31] Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, and Guiguang Ding. Box-level active detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23766–23775, 2023. 3
- [32] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36, 2024. 3
- [33] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19681–19690, 2023. 8
- [34] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Xinyu Sun, Peihao Chen, and Enming Zhang. A simple knowledge distillation framework for open-world object detection. *arXiv preprint arXiv:2312.08653*, 2023. 3, 8
- [35] Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, and Xianglong Liu. Annealing-based label-transfer learning for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11454–11463, 2023. 3
- [36] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 3
- [37] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European conference on computer vision*, pages 512–531. Springer, 2022. 8
- [38] Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788, 2024. 3
- [39] 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. 2, 3, 4
- [40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1
- [41] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 1
- [42] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the “edge” of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 3, 4, 7
- [43] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018. 3

- [44] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 2
- [45] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. 6
- [46] Leqi Shen, Tao He, Sicheng Zhao, Zhelun Shen, Yuchen Guo, Tianshi Xu, and Guiguang Ding. X-reid: Cross-instance transformer for identity-level person re-identification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1
- [47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2
- [48] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17302–17312, 2024. 3
- [49] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17302–17312, 2024. 8
- [50] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024. 2
- [51] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection, 2024. 1, 2, 5, 7
- [52] Yanghao Wang, Zhongqi Yue, Xian-Sheng Hua, and Hanwang Zhang. Random boxes are open-world object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6233–6243, 2023. 8
- [53] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 3
- [54] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023. 3
- [55] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *European Conference on Computer Vision*, pages 193–210. Springer, 2022. 3
- [56] Yizhe Xiong, Hui Chen, Tianxiang Hao, Zijia Lin, Jungong Han, Yuesong Zhang, Guoxin Wang, Yongjun Bao, and Guiguang Ding. Pyra: Parallel yielding re-activation for training-inference efficient task adaptation. In *European Conference on Computer Vision*, pages 455–473. Springer, 2025. 3
- [57] Fan Yang, Xinhao Xu, Hui Chen, Yuchen Guo, Yuwei He, Kai Ni, and Guiguang Ding. Gpro3d: Deriving 3d bbox from ground plane in monocular 3d object detection. *Neurocomputing*, 562:126894, 2023. 1
- [58] Hui-Yue Yang, Hui Chen, Ao Wang, Kai Chen, Zijia Lin, Yongliang Tang, Pengcheng Gao, Yuming Quan, Jungong Han, and Guiguang Ding. Promptable anomaly segmentation with sam through self-perception tuning. *arXiv preprint arXiv:2411.17217*, 2024. 3
- [59] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 3
- [60] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 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*, pages 23497–23506, 2023. 3
- [61] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, page 626–630. IEEE, 2022. 3
- [62] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 3
- [63] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 3
- [64] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023. 3
- [65] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 3, 4
- [66] Tiancheng Zhao, Peng Liu, Xuan He, Lu Zhang, and Kyusong Lee. Real-time transformer-based open-vocabulary detection with efficient fusion head, 2024. 7
- [67] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yixuan Qiao, Yuqing Ma, and Duorui Wang. Revisiting open world object detection, 2022. 3
- [68] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining, 2021. 3
- [69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3
- [70] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In

- ternational Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 9
- [71] Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*, 2023. 3
- [72] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2023. 8