FINDMEIFYOUCAN: BRINGING OPEN SET METRICS TO *near*, *far* AND *farther* OUT-OF-DISTRIBUTION OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, out-of-distribution (OOD) detection has gained traction as a key research area in object detection (OD), aiming to identify incorrect predictions often linked to unknown objects. In this paper, we reveal critical flaws in the current OOD-OD evaluation protocol: it fails to account for scenarios where unknown objects are ignored since the current metrics (AUROC and FPR) do not evaluate the ability to find unknown objects. Moreover, the current benchmark violates the assumption of non-overlapping objects with respect to in-distribution (ID) classes. These problems question the validity and relevance of previous evaluations. To address these shortcomings, first, we manually curate and enhance the existing benchmark with new evaluation splits—semantically *near*, *far*, and *farther* relative to ID classes. Then, we integrate established metrics from the open-set object detection (OSOD) community, which, for the first time, offer deeper insights into how well OOD-OD methods detect unknown objects, when they overlook them, and when they misclassify OOD objects as ID—key situations for reliable real-world deployment of object detectors. Our comprehensive evaluation across several OD architectures and OOD-OD methods show that the current metrics do not necessarily reflect the actual localization of unknown objects, for which OSOD metrics are necessary. Furthermore, we observe that semantically and visually similar OOD objects are easier to localize but more likely to be confused with ID objects, whereas *far* and *farther* objects are harder to localize but less prone to misclassification.

1 INTRODUCTION

In the last decade, the rise of deep learning has introduced prominent breakthroughs and achievements in object detection (OD) Zou et al. (2023), where models are usually trained under a closed-world assumption: test-time categories are the same as the training ones. However, during deployment in the real world, OD models will encounter Out-of-Distribution (OOD) objects Nitsch et al. (2021), *i.e.*, object categories different than those observed during training. While facing OOD objects, one of two safety-critical (high-risk) situations can arise: either the unknown objects are incorrectly classified as one of the In-Distribution (ID) classes, or the OOD objects will be ignored Dhamija et al. (2020).

In response to these safety challenges, researchers have developed two primary approaches: Out-of-Distribution Object Detection (OOD-OD) Du et al. (2022b) and Open-Set Object Detection (OSOD) Dhamija et al. (2020). OOD-OD focuses on identifying predictions that do not belong to the ID categories, while OSOD actively attempts to detect the unknown objects themselves. Though both approaches address the fundamental problem of encountering objects from a different semantic space than the training distribution, they employ significantly different methodologies, evaluation metrics, and benchmarks. This methodological divergence has led to isolated research communities and evaluation frameworks that fail to capture the complete picture of model performance when encountering unknown objects.

Currently, the evaluation of OOD-OD relies on a single benchmark, to the best of our knowledge: the VOS-benchmark Du et al. (2022b). The fundamental assumption of this benchmark is that none

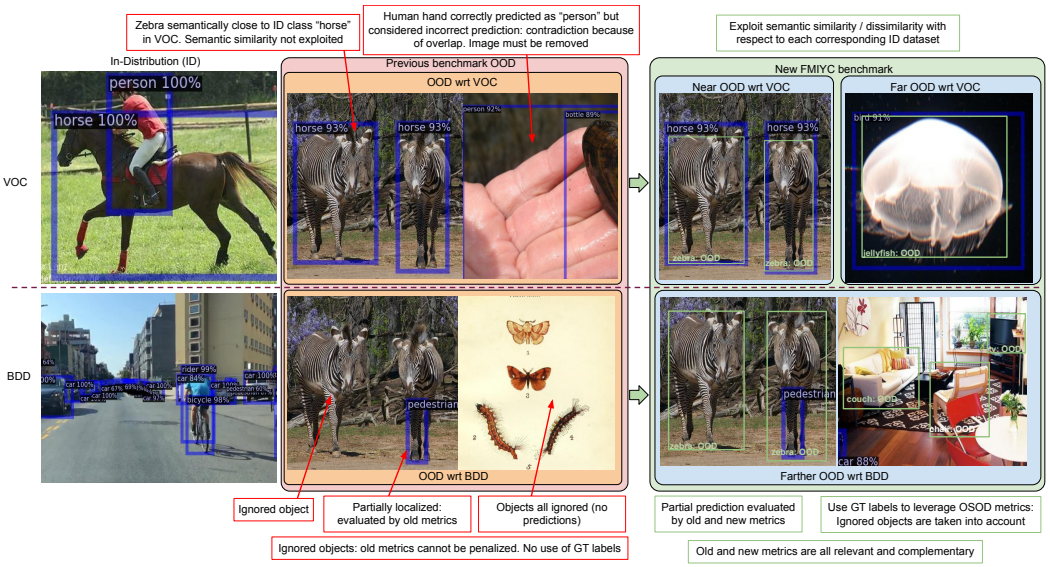

Figure 1: Predictions of Faster-RCNN trained on two ID datasets on samples from each ID and the OOD datasets in blue rectangles. The first row contains predictions of the Faster-RCNN trained on Pascal-VOC. The second row contains the predictions by the model trained on BDD100k. Ground Truth (GT) labels are shown in clear green. The base model predictions are the inputs to OOD scoring functions; without predictions, objects in images will be ignored by OOD scoring functions too. The proposed FMIYC benchmark removes undesirable semantic overlaps and separates semantically *near*, *far*, and *farther* objects with respect to the ID dataset. FMIYC uses ground truth bounding boxes to leverage OSOD metrics that measure when unknown objects are ignored, when they are detected, and when they are confounded with ID objects.

of the images in the OOD datasets include any of the ID classes, implying non-overlapping semantic spaces. Consequently, any prediction made on the OOD datasets by a model trained on the ID classes is inherently incorrect, regardless of the accuracy of object localization. The benchmark employs the area under the ROC curve (AUROC) and the false positive rate at 95% true positive rate (FPR95) as metrics. However, these metrics can be misleading, as they might suggest that a higher AUROC or lower FPR95 indicates better localization of unknown objects, which is not necessarily true. The current benchmark metrics evaluate how well OOD-OD methods identify incorrect predictions, which may potentially correspond to unknown objects. Yet, they fall short of measuring the actual identification of unknown objects. This raises a critical question: *Are AUROC and FPR95 sufficient metrics for assessing the deployment of OOD-OD methods in real-world scenarios?*

In this study, we identify and address fundamental flaws in the existing OOD-OD benchmark and its metrics, while bridging the gap between OOD-OD and OSOD research communities. We demonstrate that the current evaluation violates the fundamental assumption of non-overlap, as the OOD datasets contain ID classes. The benchmark may give the misleading impression of evaluating the identification of unknown objects, fails to penalize ignored unknown objects, and lacks proper assessment of object localization precision–issues that cannot be overlooked for safety-critical applications. To address these challenges, we propose *FindMeIfYouCan* (FMIYC), a comprehensively curated benchmark that: (1) eliminates undesired semantic overlaps between ID and OOD datasets, (2) introduces semantically stratified *near*, *far*, and *farther* OOD splits to evaluate detection robustness across varying levels of semantic similarity, and (3) properly evaluates the actual identification of unknown objects by integrating complementary metrics from the OSOD community, thus providing a robust OOD-OD evaluation framework. By combining strengths from both approaches, our benchmark enables fair comparison across multiple architectures (Faster R-CNN, YOLOv8, RT-, OWLv2) and reveals insights previously obscured in the current standard benchmark. Additionally, we adapt OOD detection methods from image classification and evaluate prominent OOD-OD methods as strong baselines for both OOD-OD and OSOD tasks, establishing a solid foundation for future research that can benefit from both perspectives.

Contributions. In summary, the main contributions of this work are:

- We identify and address fundamental flaws in the existing OOD-OD evaluation methodology, demonstrating how the current approach fails to capture a complete picture of the model's performance when encountering unknown objects.
- We propose *FindMeIfYouCan*, a benchmark that removes the existing semantic overlaps and introduces stratified *near*, *far*, and *farther* OOD splits for OOD-OD evaluation across varying levels of semantic similarity.
- We reveal the limitations of legacy AUROC and FPR95 metrics and integrate complementary metrics from the OSOD community for a comprehensive OOD-OD evaluation that captures disregarded objects.
- We assess various methods and architectures for OOD-OD. In particular, post-hoc methods from image classification, and prominent OOD-OD methods. Additionally, we expand the range of evaluated architectures, including the YOLOv8, RT-DETR, and OWLv2 architectures alongside the commonly utilized Faster R-CNN, thereby establishing robust baselines for OOD-OD.

2 Background & Related Work

2.1 Object Detection

An object detector is a model $\mathcal{M}$ that takes as input an image x and generates a bounding box $\boldsymbol{b}_i$ and classification score $\boldsymbol{c}_i$ for each i-th detected object from a predefined set of categories $\mathbb{C}$ Girshick et al. (2014). Such models are trained to localize the objects that belong to the ID classes $\mathbb{C}$ and, simultaneously, ignore the rest of the objects and the background Dhamija et al. (2020). Consequently, the object detector is usually set to function according to a given confidence threshold t^* that corresponds to the one that maximizes the mAP with respect to the ID test dataset. All objects below such threshold t^* are discarded. The model output is the set of tuples $\mathcal{M}(x; t^*) = \{(\boldsymbol{b}_i, \boldsymbol{c}_i)\}$. In the remainder of the paper, the terms "unknown" and "OOD" objects are used interchangeably, and refer to classes that do not belong to $\mathbb{C}$. Two problems can arise during real-world deployment when the model encounters an unknown object: it can be incorrectly detected as one of the ID classes with confidence above the confidence threshold t^*, or the unknown object may be ignored. Therefore, two approaches exist in the literature to address these problems: OOD-OD and OSOD.

2.2 OOD-OD & OSOD Benchmarks

Similar to OOD detection for image classification, OOD-OD is formulated as a binary classification task, that for each detected instance $(\boldsymbol{b}_i, \boldsymbol{c}_i)$ leverages a confidence scoring function $\mathcal{G}$ with its own threshold τ to calculate a per-object score $\mathcal{G}(\boldsymbol{b}_i, \boldsymbol{c}_i)$ that can distinguish between ID and OOD detections. Du et al. (2022b) introduced a benchmark that has been adopted by subsequent works Du et al. (2022a); Wilson et al. (2023); Wu & Deng (2023). This benchmark utilizes BDD100k Yu et al. (2020) and Pascal-VOC Everingham et al. (2010) as ID datasets, along with subsets of COCO Lin et al. (2014) and Open Images Kuznetsova et al. (2020) as OOD datasets. Trained models on the ID datasets are then set to perform inference on the OOD datasets.

The proposed evaluation method is deemed consistent if it adheres to the critical condition that no ID class appears in any image within the OOD datasets. Consequently, any detection within these OOD datasets is automatically classified as "incorrect", irrespective of whether the prediction corresponds to a ground truth OOD object. Conversely, all predictions on the test ID dataset are considered "correct". By employing this approach, the binary classification metrics AUROC and the FPR95 are utilized to assess the efficacy of the OOD detection method. Specifically, these metrics evaluate how effectively $\mathcal{G}(\boldsymbol{b}_i, \boldsymbol{c}_i)$ assigns different scores to predictions coming from the ID and the OOD datasets Du et al. (2022b).

On the other hand, OSOD directly adds an *unknown* class to the object detector, along with the ID classes for the training process. It was first formalized by Dhamija et al. (2020), and their goal was to tackle the fact that "unknown objects end up being incorrectly detected as known objects, often with very high confidence". Moreover, the authors propose a benchmark and associated metrics, where the goal is to accurately detect known (ID) and unknown objects simultaneously, as measured by the metrics described in Section 4.2.

The benchmarking setup of OSOD is quite different from that of OOD-OD since, in this setting, the goal is to actively and correctly localize OOD and ID objects at the same time. Also, for OSOD, there is not one commonly accepted benchmark, but many benchmarks have appeared Ammar et al. (2024); Miller et al. (2018); Han et al. (2022); Dhamija et al. (2020). The common rule is that there is one training dataset with a given set of labeled categories of objects (usually VOC, with 20 categories Everingham et al. (2010)), and there is one or several subsets of an evaluation dataset that contains the training categories and other labeled classes, semantically different from the ID ones (usually from COCO Lin et al. (2014)).

3 Pitfalls of the Current OOD-OD Benchmark

Metrics. The current benchmark uses the AUROC and the FPR95 metrics inherited from the image classification task. A misconception that may be conveyed by these metrics is that a higher AUROC or lower FPR95 means better localization of OOD objects, which is not necessarily the case. These metrics measure how well OOD-OD methods identify incorrect predictions, which may or may not correspond to ground-truth unknown objects. Therefore, these metrics do not evaluate the correct localization of OOD objects, and cannot measure when OOD objects are ignored. Figure 2 depicts such issues. For more details on the metrics, see Section C from the Appendix.

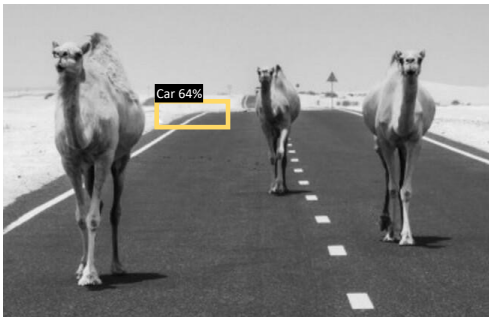

Figure 2: AUROC and FPR95 do not assess whether the relevant unknown objects, such as camels, are overlooked. They only consider incorrect predictions, such as misidentifying a car.

Semantic overlaps. The validity of previously reported results is undermined by the presence of semantic overlaps, as the OOD-OD benchmark fundamentally assumes that no ID objects appear in any OOD dataset. Under this assumption, all model predictions on OOD datasets should be considered incorrect. However, this core assumption is violated, as demonstrated in Figure 1: both labeled and unlabeled instances of people and parts of people are present in the OOD datasets. To maintain benchmark consistency, all OOD images containing ID classes must be removed. For a comprehensive list of overlapping categories in each OOD dataset and further examples, refer to Section A from the Appendix.

Ignored objects. As shown in Figure 1, not all images in each OOD dataset receive at least one prediction. Table 1 reveals that up to 59% of images in one OOD split lack any prediction above the threshold t^*. Consequently, the AUROC and FPR95 metrics reported in prior studies, such as Du et al. (2022b); Wilson et al. (2023); Du et al. (2022a); Wu & Deng (2023), are computed using only about 40% of the images in that split. By design, the benchmark's metrics are not penalized for this omission, effectively ignoring a significant portion of images and objects. To address this limitation, we advocate for the adoption of the OSOD metrics introduced in Section 4.2.

Table 1: Percentage of images with no predictions in the current OOD-OD benchmark. OI=OpenImages

Model	ID: VOC	ID: BDD
	OI/COCO	OI/COCO
F-RCNN	27.43/35.81	59.23/45.27
F-RCNN VOS	24.08/32.58	53.72/40.43

Lack of use of ground truth labels. Accurate localization of ground truth (GT) unknown objects is a critical aspect that current benchmarks overlook. A robust evaluation of a system's handling of unknown objects must go beyond simply detecting incorrect predictions. While identifying false positives is important, ignoring unknown objects can be just as risky as misclassifying them (see Figure 2). The OSOD community has established metrics to assess how well methods localize unknowns and to quantify cases where unknowns are either overlooked or confused with in-distribution (ID) objects. To further refine this evaluation, we advocate for the use of GT labels in conjunction with the OSOD metrics outlined in Section 4.2, enabling a more granular and insightful analysis.

4 THE FMIYC BENCHMARK

4.1 CREATING THE EVALUATION SPLITS

Our newly proposed FMIYC benchmark is built on top of the previous one Du et al. (2022b), by refining and enriching it in terms of overlap removal, addition of new images, splitting into subsets according to semantic similarity w.r.t. ID datasets, and the addition of open set metrics. All these factors enable fine-grained evaluation of OOD-OD. The first step involved removing overlaps. An automated process first eliminated all labeled instances of overlapping categories. Next, a manual review ensured that no unlabeled ID category instances remained in the datasets.

Then, building on established approaches in OOD detection for image classification–where OOD datasets are divided into semantically and visually *near* and *far* subsets Zhang et al. (2024); Yang et al. (2023)–we partitioned our OOD datasets w.r.t. Pascal-VOC using class names as the criterion. We matched Pascal-VOC categories (e.g., television, dog, cat, horse, cow, couch) with semantically and visually similar OOD classes (e.g., laptop, fox, bear, jaguar, leopard, cheetah, zebra, bed), assigning these to the *near* subset. All remaining OOD images, lacking a close ID counterpart, were classified as *far*. The splits were validated using WordNet Miller (1995) and the Wu-Palmer similarity metricWu & Palmer (1994), with results in Table 9 (Appendix Section B) confirming the stratification. A manual review further ensured that no near-category instances remained in the *far* subset, and vice versa. This process was applied to both COCO and OpenImages, yielding four distinct OOD subsets: COCO-near, COCO-far, OpenImages-near, and OpenImages-far. A complete list and discussion of the *near* OOD categories is available in Appendix Section A.

Table 2: Number of images in each subset of the newly proposed benchmark. CC=COCO, OI=OpenImages

ID	OOD	No. Images
VOC	CC Near	1174
	OI Near	908
	CC Far	938
	OI Far	1179
BDD	CC Farther	1873
	OI Farther	1695

We selectively incorporated additional images from the original COCO and OpenImages datasets to enrich the newly created *near* and *far* splits. The whole process was documented by recording image IDs in configuration files for each subset, ensuring full reproducibility. Both the code for generating these splits and the resulting datasets will be made publicly available.

For BDD100k as the in-distribution (ID) dataset, only overlapping images were removed, without creating separate *far* or *near* subsets or adding new images. This decision is justified by the findings in Figure 9a and Table 9, which demonstrate that BDD100k is already more distant from its respective OOD datasets than Pascal-VOC. Visual examples illustrating the semantic and visual similarity across all ID and OOD datasets are provided in Appendix Section A. These observations allow us to define three degrees of similarity between ID and OOD datasets: *near* and *far* for OOD datasets relative to Pascal-VOC, and—based on Table 9, Figure 9b, and our results—*farther* for OOD datasets relative to BDD100k. The number of images in each subset of the new benchmark is detailed in Table 2. Additionally, we assessed the similarity of each new split with respect to ID datasets in the image space using CLIP vision embeddings, as shown in Figure 3.

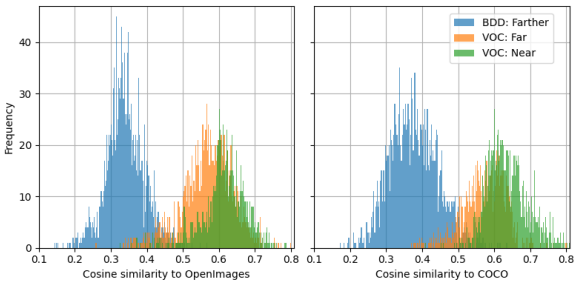

Figure 3: Perceptual and semantic (cosine) similarity Mayilvahanan et al. (2023) between ID and OOD datasets using CLIP image encoder embeddings.

4.2 PROPOSED METRICS

OSOD Metrics. The OSOD community uses as metrics the *absolute open-set error* (AOSE), the *wilderness impact* (WI), the *unknown precision* (P_U), *unknown recall* (R_U), and the *average precision of the unknowns* AP_U Gupta et al. (2022); Miller et al. (2018); Maaz et al. (2022). The AOSE reports the absolute number of unknown objects incorrectly classified as one of the ID classes. WI evaluates the proportion of AOSE among all the known detections. Unknown recall R_U is the ratio

of unknown detected objects by the number of unknown ones, and the unknown precision P_U is the ratio of true positive detections divided by all the detections Ammar et al. (2024). The OSOD metrics are fine-grained in the sense that they assess how well the methods can localize and correctly classify known and unknown objects in images where both types of objects appear.

In addition to the widely used metrics of AUROC and FPR95, we propose using the following OSOD metrics: AP_U, P_U, and R_U. We omit the WI since our benchmark does not allow both ID and OOD classes in the OOD datasets. In addition, we propose a new metric that we call *normalized open set error* (nOSE), which is the AOSE divided by the total number of labeled unknowns. We propose this metric since the absolute number of unknowns depends on the dataset, and therefore, the AOSE is not comparable across datasets, whereas the nOSE is. The nOSE assesses the proportion of unknown objects detected as one of the ID classes. A summary of the overall metrics used in the FMIYC benchmark can be found in Appendix Section C.

5 Experiments and Results

5.1 Object Detection Architectures

We used the Faster-RCNN Girshick et al. (2014) in its *vanilla* and VOS (regularized) versions, YOLOv8 Jocher et al. (2023); Sohan et al. (2024) and RT-DETR Zhao et al. (2024). As an extension, we include results from OWLv2 Minderer et al. (2024), which is a state-of-the-art VLM for object detection. For YOLOv8 and RT-DETR, the models were trained on the same ID datasets (Pascal-VOC and BDD100k). The training details can be found in Appendix Section E. For the Faster-RCNN models, we used the pre-trained checkpoints provided by Du et al. (2022b). For OWLv2, we used the original pretrained model Minderer et al. (2024). Table 3 shows the architectures mAP for each ID test dataset.

Table 3: mAP across architectures for VOC & BDD ID datasets

Model	VOC	BDD
F-RCNN	48.7	31.20
F-RCNN VOS	48.9	31.30
Yolov8	54.73	32.15
RT-DETR	70.4	33.30
OWLv2	73.2	30.40

5.2 Out-of-Distribution Object Detection Methods

We implemented prominent methods from OOD detection literature on image classification. Specifically, we selected *post-hoc* methods, as they do not require retraining of the base model. Consequently, we adapted the common families of methods from image classification to operate at the object level, as detailed below.

Output-based post-hoc methods take the logits, or the softmax activations, as inputs to their scoring functions. Here we can find MSP Hendrycks & Gimpel (2016), energy score Liu et al. (2020), and and GEN Liu et al. (2023). **Feature-space post-hoc methods** use the previous-to-last activations as the input to the scoring functions. To this category belong kNN Sun et al. (2022), DDU Mukhoti et al. (2023) and Mahalanobis Lee et al. (2018). **Mixed output-feature-space post-hoc methods** rely on the previous-to-last activations and the outputs as the input to the scoring functions. Here we find ViM Wang et al. (2022), ASH Djurisic et al. (2022), DICE Sun & Li (2022), and ReAct Sun et al. (2021). **Latent-space post-hoc methods** take inspiration from recent works Yang et al. (2023); Mukhoti et al. (2023); Arnez et al. (2024) and implement an adapted confidence score, called LaRD, that uses latent activations of a given intermediate or hidden layer.

Adapting *post-hoc* methods for object detection is straightforward, leveraging each architecture's built-in filtering mechanisms. In YOLOv8, however, only MSP, GEN, and energy-based methods are applied, as the network lacks a final fully connected layer or object-specific latent features. In addition to the adapted *post-hoc* OOD detection methods, we evaluated prominent OOD-OD methods such as VOS Du et al. (2022b), SAFE Wilson et al. (2023), and SIREN Du et al. (2022a). The confidence score threshold for each OOD detection method was calculated such that 95% of the ID samples lie above the threshold. Furthermore, as a baseline for OSOD methods in our benchmark, and to enable a fair comparison with OOD-OD methods, we present results for OpenDet CWA Mallick et al. (2024), a state-of-the-art OSOD method based on Faster-RCNN.

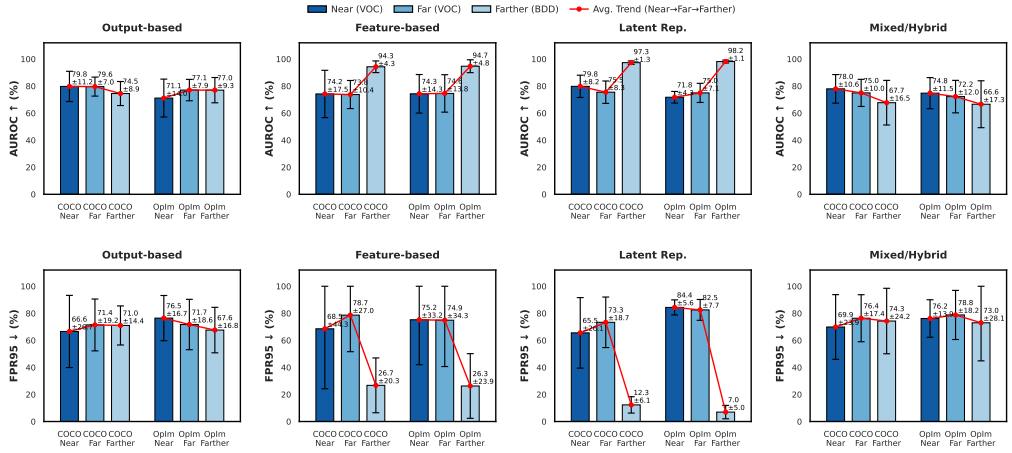

Figure 4: Average OOD-OD performance across baseline families and classic metrics (architectures are averaged).

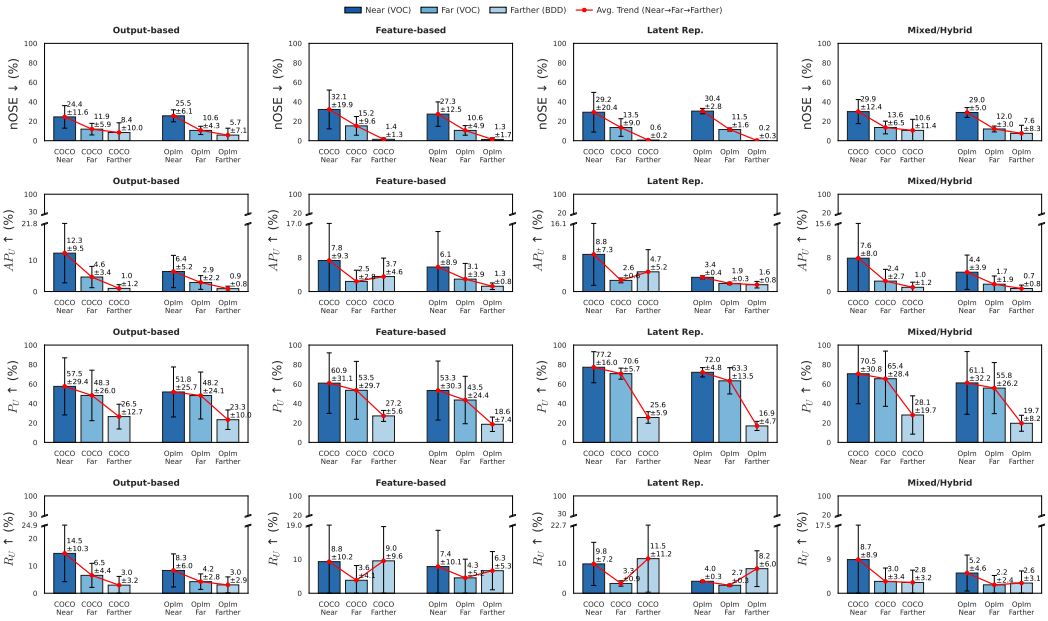

Figure 5: Average OSOD performance comparison across baseline families and metrics (architectures are averaged).

5.3 Results

In Figure 4, we present a summarized plot of the AUROC and FPR95 metrics from the new FMIYC benchmark, averaged across architectures for each family of methods and each OOD dataset. Feature-based methods and those utilizing latent representations tend to identify incorrect predictions more effectively in the *farther* split compared to other splits. Conversely, mixed methods exhibit a decline in performance as semantic distance increases. Overall, there is no distinct trend among baseline families indicating whether incorrect detections are more easily identified for *near*, *far*, or *farther* objects. This observation may be surprising; however, the differences among splits will become more apparent when considering the OSOD metrics discussed subsequently.

Figure 5 illustrates the results for the incorporated OSOD metrics, averaged across architectures for each family of methods and each OOD dataset. For the nOSE, there is a clear decreasing trend across method families when transitioning from *near* to *farther* splits. The *near* datasets exhibit the

Table 4: Results on the COCO datasets for methods using Faster-RCNN (top) and OWLv2 (bottom). **Bold:** best OOD-OD method

	AUROC ↑			R_U ↑			P_U ↑			nOSE ↓		
Method	Near	Far	Farther	Near	Far	Farther	Near	Far	Farther	Near	Far	Farther
GEN	87.43	84.48	78.82	26.12	10.96	2.99	73.80	65.17	22.89	**14.29**	**8.69**	2.04
Energy	86.47	82.31	72.44	24.84	9.95	2.99	**75.88**	**66.33**	22.89	15.95	9.80	2.03
VOS	**89.98**	**89.13**	84.79	24.62	11.26	**4.72**	72.10	55.61	**26.70**	20.49	9.65	**1.76**
SAFE	83.94	79.73	**90.73**	16.78	6.31	2.45	54.85	45.78	20.87	35.45	18.73	3.22
SIREN	89.63	88.00	-	**27.30**	**12.17**	-	60.52	53.67	-	19.46	9.84	-
OpenDet CWA	-	-	-	37.85	24.59	5.39	77.69	54.72	29.19	25.19	12.57	8.30
OWLv2 Energy	55.02	58.79	59.45	0.0	0.0	0.0	0.0	0.0	0.0	1.18	0.15	0.01
OWLv2 Mahalanobis	61.35	89.49	99.31	0.0	0.05	0.01	0.0	2.94	3.70	1.18	0.10	0.0

Table 5: Results on the OpenImages datasets for methods using Faster-RCNN (top) and OWLv2 (bottom). **Bold:** best OOD-OD method

	AUROC ↑			R_U ↑			P_U ↑			nOSE ↓		
Method	Near	Far	Farther	Near	Far	Farther	Near	Far	Farther	Near	Far	Farther
GEN	82.77	83.70	79.65	16.95	6.92	3.31	72.01	68.04	21.80	**15.86**	**5.37**	0.69
Energy	81.49	81.79	73.33	15.22	6.58	3.35	**73.59**	**70.08**	22.08	18.07	5.76	0.65
VOS	**84.40**	**86.01**	88.08	12.77	7.09	**5.63**	64.11	67.29	**26.24**	22.29	6.33	**0.63**
SAFE	85.18	83.33	**95.10**	14.9	4.31	3.18	55.70	55.38	17.17	26.86	9.36	1.36
SIREN	88.61	85.22	-	**20.88**	**6.527**	-	60.53	59.55	-	16.34	6.15	-
OpenDet CWA	-	-	-	27.51	14.11	5.93	73.42	62.08	32.93	19.67	5.56	8.59
OWLv2 Energy	56.85	59.36	48.14	0.0	0.0	0.0	0.0	0.0	0.0	6.67	0.88	0.0
OWLv2 Mahalanobis	70.84	87.67	99.55	0.68	0.17	0.0	23.28	20.58	0.0	5.98	0.71	0.0

highest nOSE, indicating that more objects are mistakenly predicted as one of the in-distribution (ID) classes among the correctly localized objects. Conversely, objects in the *farther* split are less confounded with ID objects. Regarding the AP_U, it is generally observed to be low across OOD datasets, with a trend of decreasing further in the *farther* datasets. This suggests that objects that are semantically *near* are localized more accurately. Feature-based methods and those utilizing latent space representations appear to perform better than other methods for the *farther* objects.

The P_U exhibits the highest variability across methods and also the highest values among the OSOD metrics. It is particularly elevated for the near splits. However, drops drastically for the *farther* objects, indicating that in such splits, more OOD predictions do not correspond to ground truth objects, as illustrated in Figure 2. Finally, the R_U is generally quite low across OOD datasets and methods, with a similar trend showing that objects in *far* and *farther* OOD datasets are harder to detect. The metrics reveal that, on average, most unknown objects are ignored (not found), and this challenge is even more pronounced for *far* and *farther* OOD objects. For the *near* splits, $\sim 14\%$ of unknown objects are correctly identified. This figure drops to approximately 3% in the *farther* splits for output-based and mixed methods. However, feature-based and latent representation methods seem to perform slightly better, identifying $\sim 9\%$ of the unknown objects in the *farther* splits. For a comprehensive presentation of the results for each architecture, method, and metric, please refer to Appendix Section F.

It is important to note how unrelated the previous OOD-OD benchmark metrics may seem with respect to the OSOD metrics. The AUROC and FPR95 cannot actually tell much difference between *far* and *near* datasets. This difference becomes clear in light of the OSOD metrics, which show that, contrary to the case of image classification, for object detection, the semantically and visually closer objects are easier to identify and localize. But when the unknown objects are too different from the ID ones, they will most likely be ignored by the methods and architectures evaluated. These insights are impossible to obtain using only the AUROC and FPR95.

Furthermore, Table 4 and Table 5 show summarized results for COCO/OpenImages with the most widely used architecture for OOD-OD, Faster-RCNN, across the two best post-hoc methods (GEN and Energy) according to our results, and including three OOD-OD training methods: VOS Du et al. (2022b), SAFE Wilson et al. (2023), and SIREN Du et al. (2022a). We include one OSOD method based on Faster-RCNN in order to make a fair comparison, OpenDet CWA Mallick et al. (2024). The tables show no clear winner in all OOD-OD and OSOD metrics. Across training methods, VOS presents the best AUROC performance in terms of near and far splits, and also shows the best P_U, R_U, and nOSE in the farther split. When comparing OOD-OD methods with OpenDet CWA,

it is possible to observe that it outperforms all other methods in OSOD metrics, which may not come as a surprise since it is specifically an OSOD method. It is worth clarifying that AUROC is not computable for OpenDet CWA (or OSOD methods in general), since OSOD is not a binary classification task, whereas OOD-OD is.

Finally, Table 4 and Table 5 also show the results for OWLv2 using two post-hoc OOD-OD methods. The results for OWLv2 must be understood considering that, on average, about 93% of the images in all OOD subsets do not have a single prediction, constraining the AUROC results to only around 7% of the evaluation images. This, along with the nOSE, indicates that the VLM makes many fewer incorrect predictions than in the case of Faster-RCNN, Yolov8, and RT-DETR. However, AUROC alone can be misleading. A closer look at R_U and P_U shows that OOD methods applied to OWLv2 fail to detect almost any unknown objects. While the model may internally recognize these objects, its output is strictly confined to the queried ID classes. This aligns with recent analysis by Miyai et al. (2024), which argues that VLMs require specialized OOD approaches that account for their prompt-based input and extensive semantic space.

6 Discussion

The value of OSOD metrics. We suggest caution to practitioners when relying solely on legacy metrics (AUROC and FPR95) and the former evaluation approach, as it does not take into account ignored objects or images without prediction, resulting in fewer 'valid' images for evaluation independently of the architecture for object detection. It is crucial to note that the OSOD metrics are necessary to quantify the effectiveness of OOD-OD methods in detecting actual OOD objects (AP_U and P_U) and accounting for instances when OOD objects are overlooked (R_U) or misclassified (nOSE). Unlike AUROC and FPR95, the OSOD metrics provide a more nuanced understanding by addressing confounding unknowns for ID objects, the oversight of OOD objects, and the localization of unknowns. The added value of the OSOD metrics is clearer when considering the semantic stratified splits.

***Near*, *far* and *farther* splits.** The partition of the benchmark into *near*, *far*, and *farther* proved insightful and meaningful since it details that semantic similarity plays an important role in the detection ability of different methods and architectures. It is especially insightful how the *near* OOD objects are more easily detectable than *far* and *farther* ones in the case of object detection. This is the opposite of the case of image classification, where *near* classes are considered harder than *far* ones. However, the *near* objects are also more easily confounded with ID objects, in agreement with image classification observations. Moreover, the observation that *far* and *farther* objects are more usually ignored, and therefore are hardly localizable, is demonstrated by the OSOD metrics, as only around 5% of the unknown *farther* objects are localized, as opposed to about 20% for some methods in the *near* datasets. Our work paves the way for newer detection approaches customized to specific semantic similarity requirements and provides a stronger foundation for developing OOO-OD and OSOD methods.

7 Conclusion

In this work, we identified and addressed fundamental flaws in the existing *de facto* out-of-distribution object detection (OOD-OD) evaluation benchmark and its metrics. To address these flaws, we introduced the *FindMeIfYouCan* benchmark, which builds on top of and refines the existing evaluation framework for OOD-OD. In addition, we propose incorporating open-set object detection metrics to comprehensively assess OOD-OD methods on their ability to identify unknown objects. The proposed benchmark approach offers and facilitates a holistic evaluation, measuring the detection of semantically *near*, *far*, and *farther* objects, instances where objects are overlooked, and cases where objects are misclassified as in-distribution (ID) objects. We believe our work lays a solid foundation for a more rigorous and nuanced evaluation of OOD-OD methods towards a more reliable deployment of object detectors in real-world scenarios.

REPRODUCIBILITY STATEMENT.

We include details throughout the paper that can be used to recreate the dataset and to reproduce our results. In particular, Section 4, and Section B from the Appendix. Upon acceptance, we will make publicly available the code used for dataset creation, the dataset created, and benchmark evaluation code, to ensure reproducibility and adoption of the benchmark.

REFERENCES

Hejer Ammar, Nikita Kiselov, Guillaume Lapouge, and Romaric Audigier. Open-set object detection: towards unified problem formulation and benchmarking. *arXiv preprint arXiv:2411.05564*, 2024.

Fabio Arnez, Daniel Alfonso Montoya Vasquez, Ansgar Radermacher, and François Terrier. Latent representation entropy density for distribution shift detection. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1021–1030, 2020.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35:20434–20449, 2022a.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022b.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

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*, pp. 9235–9244, 2022.

Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9591–9600, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL `https://github.com/ultralytics/ultralytics`.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5830–5840, 2021.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Thomas A Lasko, Jui G Bhagwat, Kelly H Zou, and Lucila Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of biomedical informatics*, 38 (5):404–415, 2005.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023.

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*, pp. 512–531. Springer, 2022.

Prakash Mallick, Feras Dayoub, and Jamie Sherrah. Wasserstein distance-based expansion of low-density latent regions for unknown class detection. *CoRR*, 2024.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3243–3249. IEEE, 2018.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Julia Nitsch, Masha Itkina, Ransalu Senanayake, Juan Nieto, Max Schmidt, Roland Siegwart, Mykel J Kochenderfer, and Cesar Cadena. Out-of-distribution detection for automotive perception. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2938–2943. IEEE, 2021.

Rafael Padilla, Sergio L Netto, and Eduardo AB Da Silva. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242. IEEE, 2020.

David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Mupparaju Sohan, Thotakura Sai Ram, Rami Reddy, and Ch Venkata. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pp. 529–545. Springer, 2024.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 691–708. Springer, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Samuel Wilson, Tobias Fischer, Feras Dayoub, Dimity Miller, and Niko Sünderhauf. Safe: Sensitivity-aware features for out-of-distribution object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23565–23576, 2023.

Aming Wu and Cheng Deng. Tib: Detecting unknown objects via two-stream information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Z Wu and M Palmer. Verbs semantics and lexical selection. inproceedings of the 32nd annual meeting on association for computational linguistics (pp. 133-138). In *Association for Computational Linguistics*, volume 2, 1994.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, pp. 1–16, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024.

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