Open Vocabulary Monocular 3D Object Detection

Jin Yao[1] Hao Gu[1] Xuweiyi Chen[1] Jiayun Wang[2] Zezhou Cheng[1]
[1] University of Virginia [2] California Institute of Technology
https://uva-computer-vision-lab.github.io/ovmono3d/

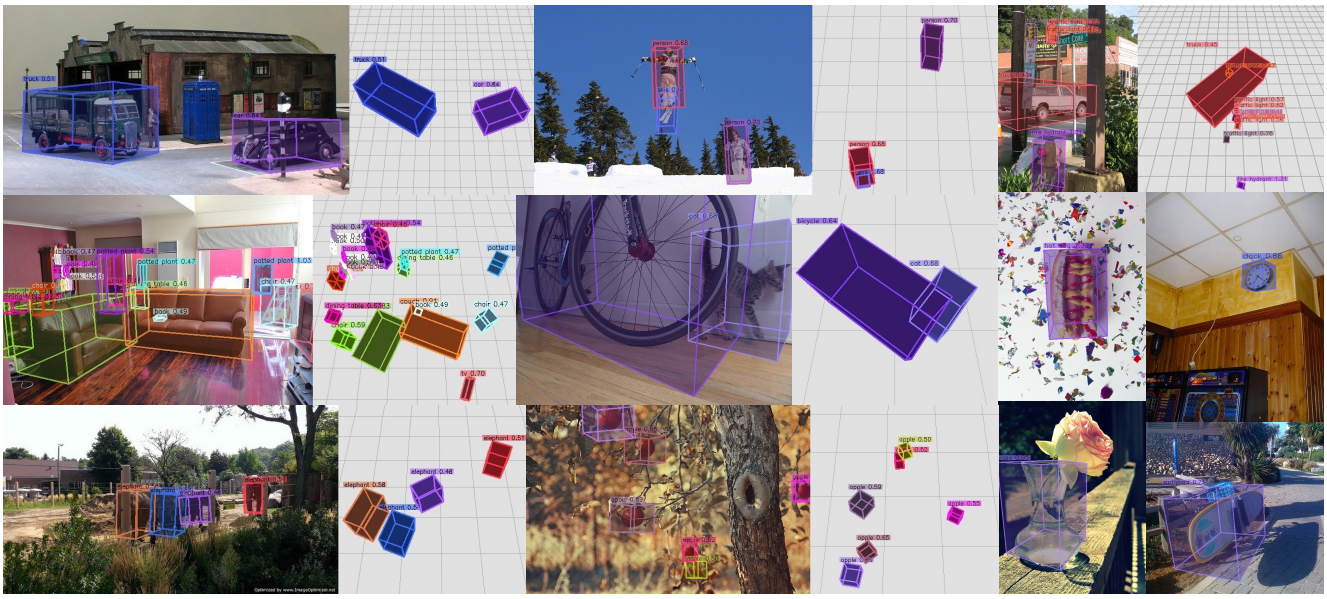


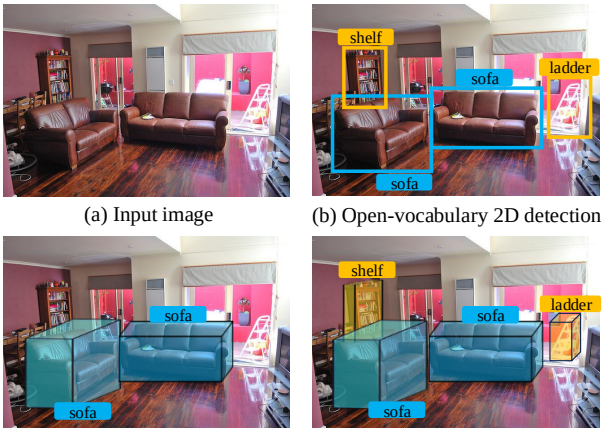
Figure 1. **Our Model on In-the-Wild COCO [35] Images**. We display 3D predictions overlaid on the images and the top-down views with a base grid of 1 m × 1 m tiles. For single-object images, only front-views are displayed. Results are shown for both base and novel categories, demonstrating that our proposed method exhibits zero-shot generalization capability on real-world images.

Abstract

*We propose and study open-vocabulary monocular 3D detection, a novel task that aims to detect objects of any categories in metric 3D space from a single RGB image. Existing 3D object detectors either rely on costly sensors such as LiDAR or multi-view setups, or remain confined to closed vocabularies settings with limited categories, restricting their applicability. We identify two key challenges in this new setting. First, the scarcity of 3D bounding box annotations limits the ability to train generalizable models. To reduce dependence on 3D supervision, we propose a framework that effectively integrates pretrained 2D and 3D vision foundation models. Second, missing labels and semantic ambiguities (*e.g.*, table vs. desk) in existing datasets hinder reliable evaluation. To address this, we design a novel metric that captures model performance while mitigating annotation issues. Our approach achieves state-of-the-art results in zero-shot 3D detection of novel categories as well as in-domain detection on seen classes. We hope our method provides a strong baseline and our evaluation protocol establishes a reliable benchmark for future research.*

1. Introduction

Recognizing objects from a single image has been a long-standing task in computer vision, with broad applications in robotics and AR/VR. Over recent decades, 2D object detection — identifying and localizing objects within a 2D image plane — has achieved substantial progress, driven by advancements in deep learning techniques [7, 16, 20, 52] and large annotated datasets [19, 35, 57]. However, recognizing only a fixed set of objects is limiting, given the vast diversity of objects in real-world settings; detecting objects solely in



(a) Input image (b) Open-vocabulary 2D detection

(c) Closed-vocabulary 3D detection (d) Open-vocabulary 3D detection

Figure 2. Given **(a)** a single image, we illustrate examples of **(b)** open-vocabulary 2D detection, which localizes objects of any category within the 2D image plane, covering both **seen** categories and **novel** categories not seen during training; **(c)** closed-vocabulary 3D detection, which detects objects from a predefined set of categories in 3D space; and **(d)** open-vocabulary 3D object detection, which identifies objects of any category in 3D.

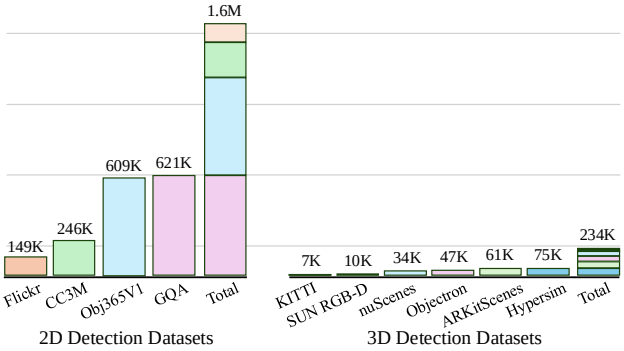


Figure 3. **2D vs. 3D Detection Datasets in terms of #Images.** Publicly available 3D datasets with 3D annotations are significantly smaller than 2D detection datasets.

2D space is also insufficient for most practical tasks, as the world and its objects exist in 3D space.

To address these limitations, one line of recent research has focused on open-vocabulary 2D object detection [10, 36, 63, 64, 76] (Fig. 2b) to identify objects beyond a fixed set of categories. Another line explores the monocular 3D detection task [3, 42, 55, 61] (Fig. 2c), which extends detection capabilities from 2D to 3D space. Despite the vast research in these two lines of research, the intersection of these two areas — open-vocabulary monocular 3D detection, referred to as OVMono3D (Fig. 2d) — remains largely unexplored. In this work, we aim to fill this gap. The OVMono3D task involves detecting and localizing objects of any categories in the metric 3D space, including novel categories unseen during training.

One of the central challenges in OVMono3D lies in **generalizable model training**, as large-scale 3D datasets with high-quality annotations are scarce. Unlike open-vocabulary 2D detection, where abundant training data exists, 3D datasets are limited in scale (Fig. 3). To address this, we adopt a decoupled strategy that separates 2D recognition and localization from 3D bounding box estimation. Specifically, we explore two complementary approaches: (1) **OVMono3D-GEO**, a simple training-free baseline that unprojects 2D detections from off-the-shelf open-vocabulary detectors [10, 36] into 3D using geometric principles (Fig. 4a). While requiring no 3D annotations and achieving reasonable performance, this approach is sensitive to occlusion and degrades in cluttered scenes; (2) **OVMono3D-LIFT**, a data-driven method that integrates multiple vision foundation models (*e.g.*, depth estimation [47], 2D open-vocabulary detection [10, 36], and general image encoder [43]) and learns to lift 2D detections into 3D (Fig. 4b). We systematically analyze its design space, including backbone architectures, 2D base detectors, and other critical choices (Sec. 5.3). Our results highlight the importance of 3D-aware image representations and accurate depth estimation for robust performance. Compared to the geometric baseline, OVMono3D-LIFT shows significantly greater resilience to occlusion and achieves superior results across diverse, real-world scenarios.

Another central challenge is **reliable model evaluation**, hindered by missing annotations and semantic ambiguity. The task requires detecting all visible objects in 2D images, yet existing 3D datasets generally provide only partial annotations due to the high cost of labeling. For example, in SUN RGB-D [59], a chair visible in the RGB image may be unannotated in 3D if it was partially occluded in the depth scan, leading to false negatives in evaluation. Similarly, inconsistencies in naming conventions across datasets, such as predicting desk when the ground truth label is table, cause correct detections to be miscounted as false positives under standard metrics. To address these issues, we propose a simple evaluation protocol that reduces the impact of missing labels and semantic variation (Sec. 4). The resulting metric enables reliable assessment without exhaustive annotations and is particularly suited for in-the-wild images, where complete labeling is infeasible and error-prone.

Finally, we conduct extensive experiments and ablations to validate both our proposed approaches and evaluation metric on standard benchmarks. Our method not only surpasses concurrent open-vocabulary 3D detectors (*e.g.*, DetAny3D [71]) by a large margin (Tab. 1), but also outperforms closed-vocabulary counterparts (*e.g.*, Cube R-CNN[3], UniMODE [34]) on in-domain categories (Tab. 2), under both standard and proposed metrics. Qualitative results further highlight strong generalization to in-the-wild

images (Fig. 1). We hope these insights inspire further exploration of advanced architectures for this task. Code will be released to support future research.

2. Related Work

Open-Vocabulary 2D Object Detection. The goal of this task is to recognize and localize objects in 2D images beyond a fixed set of predefined categories. Leveraging large-scale 2D datasets [28, 44, 57], this field has seen considerable advancement. Some approaches [18, 40, 41, 67] employ pre-trained vision-language models [27, 49], using frozen text features to detect novel categories. Other methods [10, 32, 36, 70] are pre-trained on extensive detection, grounding, and caption data to align region-text features. For example, Grounding DINO [36] incorporates grounded pre-training with cross-modal feature fusion, while YOLO-World [10] uses region-text contrastive loss and re-parameterization to enhance both accuracy and efficiency. However, such large-scale annotations are costly in 3D detection; thus, we investigate methods to adapt existing open-vocabulary 2D detectors for 3D detection.

Open-Vocabulary 3D Object Detection. This task seeks to identify objects from any category in 3D, including those unseen during training. Prior research [5, 6, 13, 38, 39, 45, 46, 66, 68, 72, 74, 80] primarily focuses on 3D detection tasks with 3D point clouds as input. [38] first proposes an open-vocabulary 3D detector using image-level class supervision from ImageNet1K [56]. [39] utilizes 2D bounding boxes from a pre-trained 2D detector [79] to build pseudo-3D label. [68] leverages various 2D foundation models to enhance the performance of 3D open-vocabulary detection. OV-Uni3DETR [62] proposes a multi-modal open-vocabulary 3D detector that accommodates both point clouds and images as the input. In contrast, our work focuses on the monocular 3D detection task that only requires RGB images as input and does not assume the availability of point cloud data at training or inference phase.

Monocular 3D Object Detection refers to the task of identifying and localizing objects within a scene using 3D bounding boxes derived from single-view images. Early research in this domain primarily targeted either outdoor [8, 9, 23, 60, 61, 75, 77, 78] or indoor [11, 25, 31, 42, 54] environments, focusing on specialized applications such as autonomous driving and room layout estimation. The extensive Omni3D [3] dataset enabled Cube R-CNN [3] and UniMODE [34] to perform unified monocular 3D object detection across multiple scene types. Recent works also explore pseudo annotation generation to expand detector vocabularies. OVM3D-Det [24] generates pseudo labels enabling detection of novel categories without human annotations. V-MIND [26] converts large-scale 2D datasets into pseudo 3D training data to expand detection vocabulary.

However, most of these approaches suffer from limited generalizability, and even advanced models [3, 34] are constrained by a closed vocabulary, restricting their ability to recognize or detect object classes that were not included during training. To address this limitation, our work focuses on exploring the potential of monocular open-vocabulary 3D detection.

Concurrent to our work, DetAny3D [71] develops a promptable 3D detection model using large-scale data and foundation models. This work addresses similar problems to ours through different methodologies. We compare with DetAny3D in our experiments quantitatively with extensive experiments.

3. Methodology

Our approach builds on two frameworks: Cube R-CNN [3] for closed-vocabulary monocular 3D detection and Grounding DINO [36] for open-vocabulary (OV) 2D detection. In this section, we first provide an overview of these frameworks (Sec. 3.1); we then introduce our proposed methods for addressing OVMono3D (Sec. 3.2 - 3.3, Fig. 4).

3.1. Preliminaries

Cube R-CNN [3] is a state-of-the-art monocular 3D detection model trained on a large-scale 3D dataset (*i.e.*, Omni3D). It extends Faster R-CNN [52] with a 3D cube head. Using 2D region proposals as input, the cube head employs ROI poolers to extract local image features and then predicts 3D bounding boxes with MLPs. The training objective of Cube R-CNN is defined as:

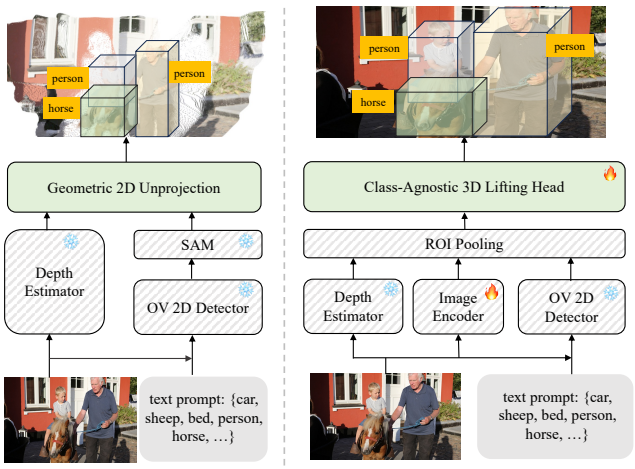
$$\mathcal{L} = \mathcal{L}_{\text{2D}} + \sqrt{2}\,\exp(-\mu)\,\mathcal{L}_{\text{3D}} + \mu, \tag{1}$$

where $\mathcal{L}_{\text{2D}}$ includes the classification and bounding box regression losses from the 2D detection head [52], $\mathcal{L}_{\text{3D}}$ is the loss from the 3D cube head, and μ denotes the uncertainty score. The 3D loss $\mathcal{L}_{\text{3D}}$ consists of disentangled losses for each 3D attribute [58]:

$$\mathcal{L}_{\text{3D}} = \sum_{a} \mathcal{L}_{\text{3D}}^{(a)} + \mathcal{L}_{\text{3D}}^{\text{all}}, \tag{2}$$

where $a \in \{(x_{\text{2D}}, y_{\text{2D}}),\ z,\ (w, h, l),\ r\}$ represents variable groups for the 2D center shift, depth, 3D dimensions (*i.e.*, width, height, and length in meters), and 3D poses. $\mathcal{L}_{\text{3D}}^{(a)}$ isolates the error of a specific group by substituting other predicted variables with ground-truth values when constructing the predicted 3D bounding box B_{3D}, while $\mathcal{L}_{\text{3D}}^{\text{all}}$ compares the vertices of predicted 3D bounding box with the ground truth using Chamfer distance.

Grounding DINO [36] is a leading framework for OV 2D detection, which combines the Transformer-based DINO



(a) OVMono3D-GEO (b) OVMono3D-LIFT

Figure 4. **Proposed Methods.** **(a)** OVMono3D-GEO is a training-free method that predicts 3D detections from 2D via geometric unprojection using off-the-shelf depth estimation (*i.e.* UniDepthv2 [48]), segmentation (*i.e.* SAM [29]), and OV 2D detector [36]. **(b)** OVMono3D-LIFT is a learning-based approach that trains a class-agnostic neural network to lift 2D detections and geometric information to 3D. Both approaches decouple 2D recognition and localization from 3D bounding box estimation.

detector [69] with grounded pretraining. We adopt a pretrained Grounding DINO as our default OV 2D detection model due to its superior performance and strong zero-shot generalization capabilities. An ablation study of different base 2D detectors on our approach is provided in Sec. 5.

3.2. Training-free OVMono3D-GEO

We first establish a training-free baseline that unprojects 2D detections into 3D using geometric principles (Fig. 4a). Given an input image I, a text prompt T, and 2D bounding boxes with category labels $\{(b_i, c_i)\}_{i=1}^{N}$ predicted by an OV 2D detector [36], the method proceeds as follows.

3D Geometric Prediction. For each detected object (b_i, c_i), an instance segmentation mask S_i is obtained using a segmentation model (*e.g.*, SAM [29]), and a metric depth estimation model (*e.g.*, UniDepthv2 [48]) generates the depth map $D \in \mathbb{R}^{H\times W}$. Pixels within S_i are unprojected into 3D space using the camera intrinsic matrix $K \in \mathbb{R}^{3\times 3}$ to form a point cloud P_i. Here, u and v denote the pixel coordinates in the image, and each 3D point $p = [x, y, z]^\top \in P_i$ is computed as:

$$p = D(u, v) \cdot K^{-1} \begin{bmatrix} u & v & 1 \end{bmatrix}^T .$$

3D Bounding Box Generation. To estimate the 3D bounding box parameters $B_i = (t_i, d_i, r_i)$ from the point cloud P_i, we first apply Principal Component Analysis (PCA) on P_i to estimate the object orientation r_i from the principal component direction. PCA determines the object's main orientation by identifying the directions of maximum variance, enabling tight bounding box fitting. DBSCAN clustering [12] is then applied to remove outliers from noisy depth predictions and imperfect segmentation masks, ensuring robust parameter estimation. Finally, the centroid t_i and dimensions d_i are computed from the cleaned point cloud along the estimated orientation r_i.

While this method does not require 3D supervision, its accuracy relies heavily on the quality of depth estimation and segmentation, and it is particularly sensitive to occlusions, where incomplete point clouds often lead to inaccurate bounding box reconstruction (see Fig. 11 in the supplementary). This approach also shares similarities with the 3D bounding box generation pipeline in OVM3D-Det [24]. However, their work employs the generated 3D boxes as pseudo labels to expand the vocabulary of training data, without directly evaluating the accuracy of the 3D box generation itself as a OVMono3D method.

3.3. Data-driven OVMono3D-LIFT

To overcome the limitations of the geometric-based approach, we propose OVMono3D-LIFT (Fig. 4b), which learns to estimate 3D bounding boxes from large-scale 3D annotations. This framework decouples OVMono3D into two stages: (1) recognizing and localizing objects in 2D with off-the-shelf open-vocabulary detectors, and (2) class-agnostically lifting the 2D bounding boxes into 3D cuboids. We detail the key components below.

Image Encoder. OVMono3D-LIFT employs a pretrained DINOv2 [43] to extract image features, and a Feature Pyramid module [33] to generates multi-scale hierarchical feature pyramid. The predicted 2D bounding boxes are used as regions of interest for ROI pooling, extracting local object features of size 7×7. These features are fed into a cube head, as described in Sec. 3.1, to predict 3D attributes.

Depth Estimator. Similar to OVMono3D-GEO, the method first obtain the predicted depth map $D \in \mathbb{R}^{H\times W}$, then unproject it to a point map $P \in \mathbb{R}^{3\times H\times W}$ using the camera intrinsic matrix $K \in \mathbb{R}^{3\times 3}$. ROI pooling is then performed on the point map using the 2D bounding boxes to extract local geometric information of size $3 \times 7 \times 7$. This geometric information is concatenated with the visual ROI features to form geometry-informed features, which are input into the 3D cube head to predict 3D box attributes. During both training and inference, the method uses pseudo depth maps obtained from a pretrained metric depth estimator—no ground truth depth maps are required. By fusing geometric information into the model, the network can better infer object depth and dimensions, producing more accurate predictions.

Class-agnostic 3D Lifting head. Unlike Cube R-CNN [3],

Methods	Novel Categories in Omni3D						Novel Dataset
	AP_{3D}↑	AP^{15}_{3D}↑	AP^{25}_{3D}↑	AP^{50}_{3D}↑	AP^{easy}_{3D}↑	AP^{hard}_{3D}↑	AP^{city}_{3D} ↑
OVM3D-Det [24]	6.74 / 18.30	11.35 / 29.69	5.95 / 15.87	0.73 / 2.31	9.70 / 22.99	5.05 / 15.63	0.04 / 0.05
DetAny3D [71]	9.22 / 21.44	-	-	-	-	-	11.05 / 15.71
OVMono3D-GEO	8.48 / 20.63	14.07 / 35.26	8.22 / 18.26	0.49 / 1.01	9.71 / 21.20	7.78 / 20.30	0.73 / 0.90
OVMono3D-LIFT*	7.31 / 18.61	11.47 / 27.95	7.06 / 17.62	0.93 / 3.48	11.39 / 26.14	4.97 / 14.31	6.57 / 10.21
OVMono3D-LIFT	**9.61 / 24.08**	**15.00 / 37.45**	**9.39 / 23.24**	**1.41 / 4.00**	**14.71 / 30.13**	**6.7 / 20.62**	**12.88 / 16.76**

Table 1. **Performance on Novel Categories and Datasets.** We report AP_{3D} on Omni3D novel categories and novel datasets, including results under IoU_{3D} thresholds of 0.15, 0.25, and 0.5, and across easy and hard categories splits for novel categories in Omni3D (Sec. 5.1). All metrics are reported in the format of standard metric / target-aware metric (Sec. 4). We notice that with UniDepthv2 [48], performance drops significantly on CityScapes3D [14]; therefore, we use Metric3Dv2 [22] for all methods requiring pseudo depth input on this dataset only. **Bold** indicates best results and "-" denotes unreported results in previous works.

Method	AP^{kit}_{3D}↑	AP^{nus}_{3D}↑	AP^{sun}_{3D}↑	AP^{hyp}_{3D}↑	AP^{ark}_{3D}↑	AP^{obj}_{3D}↑	AP^{25}_{3D}↑	AP^{50}_{3D}↑	AP^{near}_{3D}↑	AP^{med}_{3D}↑	AP^{far}_{3D}↑	AP^{all}_{3D}↑
Closed-vocabulary methods												
SMOKE [37]	-	-	-	-	-	-	-	-	-	-	-	9.60
FCOS3D [60]	-	-	-	-	-	-	-	-	-	-	-	9.80
PGD [61]	-	-	-	-	-	-	-	-	-	-	-	11.20
ImVoxelNet [55]	-	-	-	-	-	-	-	-	-	-	-	9.40
Cube R-CNN [3]	**32.47**	29.98	15.23	7.39	41.72	53.85	25.51	9.59	28.32	12.05	8.50	23.68
UniMODE [34]	29.20	**36.00**	**23.00**	**8.10**	**48.00**	**66.10**	**30.20**	**10.60**	**31.10**	**14.90**	**8.70**	**28.20**
Open-vocabulary methods												
OVM3D-Det [24]	7.59	14.25	11.82	5.34	25.23	1.93	9.33	1.36	10.51	4.82	3.71	9.49
DetAny3D [71]	**31.61**	30.97	18.96	7.17	46.13	54.42	-	-	-	-	-	24.92
OVMono3D-GEO	1.22	3.23	9.99	3.98	21.03	9.22	5.98	0.23	9.94	1.59	0.15	6.64
OVMono3D-LIFT*	25.58	30.56	19.52	9.92	49.34	62.72	29.04	11.91	33.96	12.20	7.21	27.09
OVMono3D-LIFT	31.44	**32.47**	**23.24**	**11.89**	**54.21**	**63.48**	**32.23**	**12.90**	**37.51**	**13.35**	**8.96**	**29.63**

Table 2. **Performance on Base Categories.** We report the AP_{3D} on six subsets of the Omni3D test set, including results under IoU_{3D} thresholds of 0.25 and 0.5, and across different distances (near, medium, far). AP^{all}_{3D} denotes the overall score averaged across all subsets. For open-vocabulary methods without 2D heads, we use 2D box predictions from Cube R-CNN as input. All the metrics are in standard evaluation. **Bold** indicates best results, underlined indicates second best, and "-" denotes unreported results in previous works.

our 3D cube prediction head is *class-agnostic*, while their method employs class-specific layers and per-class average size statistics, which limit generalization to novel categories. The model trains only on base categories with the training objective identical to Cube R-CNN, as described in Sec. 3.1. Since the method also train a 2D head on base categories, the model preserves detection ability on these categories. During inference, for novel categories, an OV 2D detector (*e.g.*, Grounding DINO [36]) is used to obtain 2D boxes, which serve as input to the ROI pooling module.

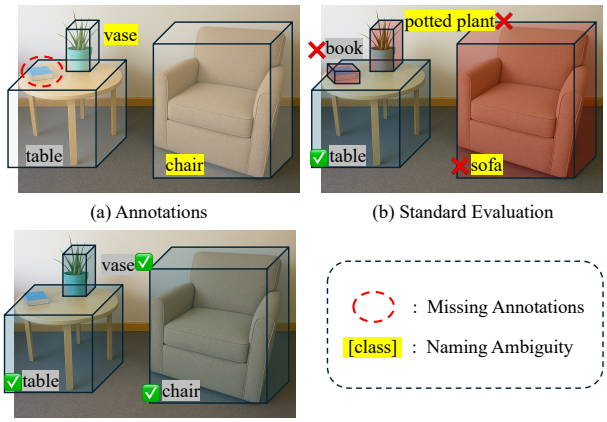
OVMono3D-LIFT*. We include a model variant as a baseline. It removes the depth estimation module from the full OVMono3D-LIFT framework, relying solely on visual features from the image encoder and Feature Pyramid Network for 3D bounding box prediction. This variant serves to evaluate the contribution of geometric depth information by comparing performance against the full model that incorporates both visual and geometric features.

4. Evaluation Metrics

Standard Metrics. Mean Average Precision (AP_{3D}) based on Intersection over Union (IoU) is the widely used standard metric for closed-vocabulary 3D object detection tasks [3, 34]. However, directly applying traditional evaluation protocols to OVMono3D presents challenges due to common issues in 3D dataset annotations:

Missing annotations: Comprehensive 3D annotation is often impractical due to high costs. Fig. 5a illustrates an example of this issue, where the book is not labeled.

Naming ambiguity: Objects may be labeled with different naming conventions or annotation policies during annotation (*e.g.*, table *vs.* desk). Standard open-vocabulary 2D detection methods typically prompt 2D detectors with exhaustive lists of possible categories, which can lead to correct predictions with class names that do not align with the dataset annotations, especially for vaguely defined or overlapping categories, as shown in Fig. 5 (chair *vs.* sofa, vase *vs.* potted plant).



(a) Annotations (b) Standard Evaluation

(c) Target-aware Evaluation

Figure 5. By prompting only categories that exist in the annotations, our target-aware evaluation mitigates the negative impact of missing annotations (*e.g.*, "book" in (a)) and naming ambiguity (*e.g.*, "vase" *vs.* "potted plant" and "chair" *vs.* "sofa".)

Target-Aware Metrics. To address these issues, we propose a simple evaluation approach that considers only the categories with ground-truth labeled instances in each image. Specifically, instead of providing the 2D detector with an exhaustive list of possible categories, we only prompt it with category names that exist in the annotations for each image. Since human annotators typically label all instances of the same category within a single image, categories with ground-truth annotations are likely to be fully annotated. Fig. 5 illustrates our approach enables more accurate model assessment despite annotation issues. For instance, the example highlights missing annotations for "books," and naming ambiguities between "chair" and "sofa." This validates the practicability of our evaluation under annotation issues.

5. Experiments

5.1. Experimental Setup

Datasets. The experiments are conducted using the Omni3D [3] dataset, the largest dataset for monocular 3D object detection across both indoor and outdoor scenes. Omni3D is repurposed and combined from six established sources including SUN RGB-D [59], ARKitScenes [2], Objectron [1], Hypersim [53], nuScenes [4], and KITTI [15]. It comprises a substantial 234k images, 3 million labeled 3D bounding boxes, and covers 98 distinct object categories.

During training, we use the 50 categories Cube R-CNN [3] trained on as the base categories. For evaluation, we select 22 categories from the remaining classes as novel categories. These categories are chosen based on two criteria: the number of test instances and the precision of category naming. To facilitate a detailed assessment of zero-shot generalization capabilities, we further divide these categories into *easy* and *hard* subsets according to object visibility. See Appendix B for detailed splits.

Besides evaluating novel categories in Omni3D [3], we perform zero-shot evaluation on a novel dataset with novel camera models. CityScapes3D [14] includes base categories from the self-driving domain. We use the same evaluation split as DetAny3D [71].

Baselines. For novel categories, we compare against OVM3D-Det [24], which generates pseudo 3D labels by combining Grounding DINO [36], SAM [29], UniDepth [47], and LLM-generated size priors. We evaluate their labeling pipeline as a baseline detector against ground truth annotations. We also include comparisons with DetAny3D [71]. For base categories, we compare against closed-vocabulary methods including SMOKE [37], FCOS3D [60], PGD [61], ImVoxelNet [55], Cube R-CNN [3], and UniMODE [34], as well as open-vocabulary methods including OVM3D-Det [24] and DetAny3D [71].

Evaluation Metrics. We report the mean Average Precision in 3D (AP_{3D}) computed across all evaluated categories. Following the Omni3D [3] evaluation protocol, predictions are matched to ground truth using 3D Intersection over Union (IoU_{3D}), which measures the volumetric overlap between predicted and ground truth 3D bounding boxes. The evaluation is conducted across IoU_{3D} thresholds $\tau \in [0.05, 0.10, ..., 0.50]$, with the final AP_{3D} representing the mean average precision across all thresholds and categories. For evaluation on novel categories and novel datasets, we additionally report target-aware metrics (see Sec. 4).

Implementation Details. Our implementation is based on PyTorch3D [51] and Detectron2 [65]. We initialize the image feature encoder with pre-trained DINOv2 base weights [43]. The model is trained on eight NVIDIA A100 GPUs for five days. For reference, DetAny3D [71] is trained on 64 A100 GPUs for two weeks. The models are trained for 116k steps with a batch size of 64 using SGD optimizer with an initial learning rate of 0.0012, which decays by a factor of 10 at 60% and 80% of training. We apply standard image augmentations including random horizontal flipping and resizing during training.

Due to computational constraints, models trained in the analysis section (Sec. 5.3) employ a resource-efficient setting where the pre-trained image encoder is kept frozen during training and the depth estimator is excluded. Unless explicitly stated otherwise, all methods use Grounding DINO [36] for 2D detection on novel categories, and all methods requiring pseudo depth input use UniDepthv2 [48].

5.2. Model Performance

Fig. 6 shows qualitative results on the Omni3D test set. In comparison with Cube R-CNN [3], OVMono3D-LIFT de-

Figure 6. **Qualitative Visualizations on the Omni3D Test Set.** For each example, we present the ground truth annotations, the predictions of Cube R-CNN and OVMONO3D-LIFT, displaying the 3D predictions overlaid on the image. For OVMONO3D-LIFT, we also present a top-down view with a base grid of 1 m $\times$ 1 m tiles. Base categories are depicted with brown cubes, while novel categories are represented in other colors. Zoom in for best viewing. See Appendix C for more visualizations.

tects not only objects of base categories, but also novel categories that are unseen at the training time.

Novel Category and Dataset Performance. Tab. 1 shows that OVMONO3D-LIFT achieves state-of-the-art performance on novel categories and novel datasets.

Geometry-based methods OVM3D-Det [24] and OVMONO3D-GEO show lower performance, particularly on Cityscapes3D [14], due to their sensitivity to occlusions and noisy depth estimation. The performance improvement from LIFT* to LIFT underscores the importance of incorporating additional geometric information from foundation depth estimators into the model. This enhanced performance demonstrates that our geometry-informed 2D lifting design effectively exploits 2D data-driven priors, including OV 2D detectors, self-supervised features, and metric depth estimators. By decomposing the task into OV 2D detection and 2D-to-3D lifting, our approach addresses challenges associated with low-quality and limited 3D annotations.

Base Category Performance. Tab. 2 compares OVMONO3D-LIFT with baselines on base categories. OVMONO3D-LIFT achieves the best performance among both closed-vocabulary and open-vocabulary methods, with an overall AP_{3D} of 29.63. It surpasses Cube R-CNN [3] by 5.95 points and UniMODE [34] by 1.43 points, while additionally providing generalization to novel categories. Compared to the open-vocabulary baseline DetAny3D [71], it achieves a 4.71-point improvement, demonstrating strong performance on base categories. Notably, OVMONO3D-LIFT* also achieves competitive results with 27.09 overall AP_{3D}, ranking second among open-vocabulary methods. These results validate the effectiveness of our approach.

Methods	$IoU_{3D}\uparrow$	XY$\uparrow$	Depth$\uparrow$	Size$\uparrow$	Pose$\uparrow$
OVMONO3D-GEO	18.69	47.96	30.97	18.68	37.84
OVMONO3D-LIFT*	18.11	49.39	27.23	48.10	72.44
OVMONO3D-LIFT	**22.03**	**52.91**	**32.59**	**49.64**	**73.57**

Table 3. **Disentangled Metrics on Novel Categories.** We report the overall and disentangled IoU_{3D} (%) for different attribute predictions on Omni3D's novel categories. For fair comparison, same 2D ground truth box inputs are applied for all methods.

5.3. Analysis

Disentangled Metrics. To analyze the contributions of different predicted attributes to 3D bounding box errors, we report disentangled IoU_{3D} for position (XY), depth, dimensions, and pose. For each attribute, we compute the IoU_{3D} of a modified 3D bounding box where only that attribute is predicted, while all others are set to their ground-truth values. The IoU_{3D} is then computed against the ground-truth cube. To facilitate the comparison, for all methods we use ground truth 2D box as input.

Tab. 3 exhibits the overall and disentangled IoU_{3D} for different attribute predictions. For LIFT* and LIFT methods, object depth prediction consistently contributes the most error to the overall prediction. This indicates that object depth estimation is the primary bottleneck of our task. Furthermore, the size and pose predictions in GEO method exhibit larger errors, indicating that geometry-based methods are not effective for these predictions. This underscores

2D Box Input	AP_{2D}↑	AP_{3D}↑	AP^{15}_{3D}↑	AP^{25}_{3D}↑	AP^{50}_{3D}↑
YOLO-World [10]	19.99	20.31	31.50	19.42	2.85
Grounding DINO [36]	**21.44**	**24.08**	**37.45**	**23.24**	**4.00**

Table 4. **Ablation on 2D Bounding Box Input.** Evaluation of open-vocabulary 2D detectors and their impact on 2D and 3D detection performance for novel categories in Omni3D.

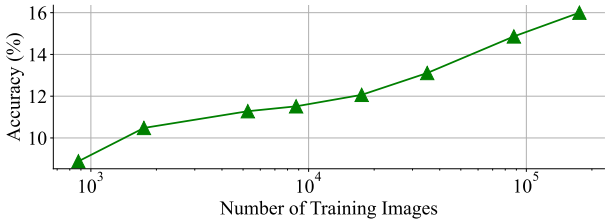


Figure 7. The performance of OVMono3D-LIFT as a function of training data amount. We report the AP_{3D} (%) on novel categories as the evaluation metric.

the necessity of developing learning-based methods.

2D Bounding Box Input. Tab. 4 evaluates the impact of 2D detectors on both 2D and 3D detection performance for novel categories. We test two state-of-the-art 2D detectors: YOLO-World [10] and Grounding DINO [36]. Grounding DINO consistently outperforms YOLO-World across all evaluation metrics, demonstrating that Grounding DINO is a preferable off-the-shelf 2D detector for OVMono3D.

Training Data Scaling Law. Fig. 7 reports our model's AP_{3D} score as a function of the training data size. This underscores the critical importance of dataset size in OVMono3D tasks and suggests that our model may achieve even better performance with more extensive training data.

Pre-trained feature extractor Selection. Tab. 5 shows the impact of pretrained feature extractors on 3D detection performance for novel categories. DINOv2 achieves the best performance across all evaluation metrics. This result underscores the effectiveness of DINOv2's representations for the 3D detection task. Our findings align with recent studies that highlight DINOv2's strong zero-shot capabilities in understanding depth, multi-view correspondences, and relative pose [17, 30, 73], indicating that DINOv2's 3D-aware image features are highly suitable for this task.

For a more detailed analysis on the role of synthetic data and quantitative evaluation of naming ambiguity issues in current benchmarks, refer to Appendix E.

5.4. Zero-Shot Generalization Performance

Fig. 1 presents OVMono3D-LIFT prediction on in-the-wild images from COCO [35]. The results show 2D projections of our predictions are well-aligned with the objects, and their top-down views closely match the visual scene layout. Even on completely out-of-distribution categories such as elephant and apple, our method produces promising results. This suggests that our model demonstrates zero-shot generalization ability on real-world images. See Fig. 10 in Supplementary Material for more qualitative results on COCO.

Feat. Extractor	Supervision	AP_{3D}↑	AP^{15}_{3D}↑	AP^{25}_{3D}↑	AP^{50}_{3D}↑
MAE [21]	SSL	7.72	11.95	6.43	0.80
CLIP [49]	VLM	9.02	14.19	8.44	0.50
MiDas [50]	Depth	10.65	17.42	9.87	0.69
DINOv2 [43]	SSL	**16.04**	**24.57**	**14.97**	**2.21**

Table 5. **Ablation Study on Feature Extractors for OVMono3D-LIFT in 3D Detection on Novel Categories.** We report AP_{3D} for various feature extractors with frozen parameters during training.

6. Discussion

We investigate OVMono3D, an under-explored task of recognizing and localizing objects from any categories in 3D using a single image. We identify unique challenges, notably data scarcity and limitations of standard evaluation metrics. We propose simple yet effective approaches, including geometry-based OVMono3D-GEO and learning-based OVMono3D-LIFT — both decouple 2D detection from 3D bounding box prediction. Such design enables the full utilization of off-the-shelf open-vocabulary 2D detectors pre-trained on large-scale datasets. Our analysis pinpoints the key components in the framework, including 3D-aware image features, base 2D detectors, and the impact of dataset scaling. We further demonstrate the zero-shot generalizability of our approach on in-the-wild images. Our findings suggest that dataset scale and accurate depth perception remain the major bottlenecks in this task. One promising direction may be to develop unsupervised learning that harness the abundance of unlabeled images. We hope this work inspires future research toward advancing this task. See Supplementary Material A for limitations.

7. Acknowledgement

The authors acknowledge the University of Virginia Research Computing and Data Analytics Center, Advanced Micro Devices AI and HPC Cluster Program, Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, and National Artificial Intelligence Research Resource (NAIRR) Pilot for computational resources, including the Anvil supercomputer (National Science Foundation award OAC 2005632) at Purdue University and the Delta and DeltaAI advanced computing resources (National Science Foundation award OAC 2005572). This work was partly supported by Adobe Research Gift.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 6

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS*, 2021. 6

[3] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023. 2, 3, 4, 5, 6, 7, 1

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 6

[5] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *NeurIPS*, 2024. 3

[6] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection. *arXiv preprint arXiv:2406.00830*, 2024. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[8] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation, 2021. 3

[9] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 3

[10] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, 2024. 2, 3, 8

[11] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *CVPR*, 2016. 3

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 4

[13] Djamahl Etchegaray, Zi Huang, Tatsuya Harada, and Yadan Luo. Find n' propagate: Open-vocabulary 3d object detection in urban environments. In *ECCV*, 2024. 3

[14] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020. 5, 6, 7

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6, 2

[16] R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 1

[17] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *ECCV*, 2022. 8

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 8

[22] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5

[23] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, pages 4012–4021, 2022. 3

[24] Rui Huang, Henry Zheng, Yan Wang, Zhuofan Xia, Marco Pavone, and Gao Huang. Training an open-vocabulary monocular 3d detection model without 3d data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 4, 5, 6, 7

[25] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation, 2019. 3

[26] Jin-Cheng Jhang, Tao Tu, Fu-En Wang, Ke Zhang, Min Sun, and Cheng-Hao Kuo. V-mind: Building versatile monocular indoor 3d detector with diverse 2d annotations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9577–9586. IEEE, 2025. 3

[27] 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 *ICML*, pages 4904–4916, 2021. 3

[28] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 3

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4, 6

[30] Akshay Krishnan, Abhijit Kundu, Kevis-Kokitsi Maninis, James Hays, and Matthew Brown. Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In *ECCV*, 2025. 8

[31] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. *arXiv preprint arXiv:2412.04458*, 2024. 3

[32] 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 *CVPR*, 2022. 3

[33] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 4

[34] Zhuoling Li, Xiaogang Xu, SerNam Lim, and Hengshuang Zhao. Towards unified 3d object detection via algorithm and data unification, 2024. 2, 3, 5, 6, 7

[35] 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 *ECCV*, 2014. 1, 8, 4, 5

[36] 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. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4, 5, 6, 8, 1

[37] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 996–997, 2020. 5, 6

[38] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. *arXiv preprint arXiv:2207.01987*, 2022. 3

[39] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, 2023. 3

[40] 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 *CVPR*, 2022. 3

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 3

[42] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image, 2020. 2, 3

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 6, 8, 1

[44] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 2011. 3

[45] Xingyu Peng, Yan Bai, Chen Gao, Lirong Yang, Fei Xia, Beipeng Mu, Xiaofei Wang, and Si Liu. Global-local collaborative inference with llm for lidar-based open-vocabulary detection. In *ECCV*. Springer, 2025. 3

[46] Xingyu Peng, Si Liu, Chen Gao, Yan Bai, Beipeng Mu, Xiaofei Wang, and Huaxia Xia. Glrd: Global-local collaborative reason and debate with psl for 3d open-vocabulary detection. *arXiv preprint arXiv:2503.20682*, 2025. 3

[47] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 2, 6

[48] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 4, 5, 6

[49] 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 *ICML*, 2021. 3, 8

[50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020. 8

[51] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 6

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016. 1, 3

[53] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 6, 3

[54] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection, 2021. 3

[55] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 2, 5, 6

[56] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[57] 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 *ICCV*, 2019. 1, 3

[58] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection. In *ICCV*, 2019. 3

[59] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 2, 6, 3, 4, 5

[60] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection, 2021. 3, 5, 6

[61] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2, 3, 5, 6

[62] Zhenyu Wang, Yali Li, Taichi Liu, Hengshuang Zhao, and Shengjin Wang. Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. *arXiv preprint arXiv:2403.19580*, 2024. 3

[63] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *T-PAMI*, 2024. 2

[64] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *AAAI*, pages 6117–6125, 2024. 2

[65] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[66] Zhongyu Xia, Jishuo Li, Zhiwei Lin, Xinhao Wang, Yongtao Wang, and Ming-Hsuan Yang. Openad: Open-world autonomous driving benchmark for 3d object detection. *arXiv preprint arXiv:2411.17761*, 2024. 3

[67] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[68] Dongmei Zhang, Chang Li, Renrui Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang Zhang. Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. In *AAAI*, 2024. 3

[69] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 4

[70] 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. *NeurIPS*, 2022. 3

[71] Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild. *arXiv preprint arXiv:2504.07958*, 2025. 2, 3, 5, 6, 7

[72] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. Opensight: A simple open-vocabulary framework for lidar-based object detection. In *ECCV*. Springer, 2025. 3

[73] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *NeurIPS*, 2024. 8

[74] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 3

[75] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, pages 9155–9166, 2023. 3

[76] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. 2

[77] Dingfu Zhou, Xibin Song, Yuchao Dai, Junbo Yin, Feixiang Lu, Jin Fang, Miao Liao, and Liangjun Zhang. Iafa: Instance-aware feature aggregation for 3d object detection from a single image, 2021. 3

[78] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019. 3

[79] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 3

[80] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023. 3