Just Add Geometry: Gradient-Free Open-Vocabulary 3D Detection Without Human-in-the-Loop

Atharv Goel Mehar Khurana
{atharv21027, mehar21541}@iiitd.ac.in

Indraprastha Institute of Information Technology, Delhi

Abstract

Modern 3D object detection datasets are constrained by narrow class taxonomies and costly manual annotations, limiting their ability to scale to open-world settings. In contrast, 2D vision-language models trained on web-scale image-text pairs exhibit rich semantic understanding and support open-vocabulary detection via natural language prompts. In this work, we leverage the maturity and category diversity of 2D foundation models to perform open-vocabulary 3D object detection without any human-annotated 3D labels.

Our pipeline uses a 2D vision-language detector to generate text-conditioned proposals, which are segmented with SAM and back-projected into 3D using camera geometry and either LiDAR or monocular pseudo-depth. We introduce a geometric inflation strategy based on DBSCAN clustering and Rotating Calipers to infer 3D bounding boxes without training. To simulate adverse real-world conditions, we construct ***Pseudo-nuScenes****, a fog-augmented, RGB-only variant of the nuScenes dataset.*

Experiments demonstrate that our method achieves competitive localization performance across multiple settings, including LiDAR-based and purely RGB-D inputs, all while remaining training-free and open-vocabulary. Our results highlight the untapped potential of 2D foundation models for scalable 3D perception. We open-source our code and resources at https://github.com/atharv0goel/open-world-3D-det.

1. Introduction

Object detection is a foundational task in computer vision, critical for applications such as autonomous driving [2], augmented reality, and robotics [22]. While traditional approaches focus on detecting objects in 2D images [17, 18], the need for richer spatial understanding has led to the development of 3D object detection methods that operate on modalities like point clouds or RGB-D data [7, 16]. These methods reason about object geometry and spatial layout in three-dimensional space, offering more precise localization and scene comprehension.

Despite this progress, 3D object detection remains constrained by the limited class taxonomies available in existing 3D datasets, which typically contain only a handful of object categories. For instance, KITTI [4] includes just three classes, while nuScenes [2] and Waymo [19] offer 10–23. In contrast, 2D detection datasets such as COCO [9] or LVIS [5] encompass hundreds or thousands of categories, providing a much broader object vocabulary. This discrepancy hampers the generalization capability of 3D detectors, particularly in open-world or real-world scenarios where object classes are diverse, fine-grained, and continually evolving [24].

Open-vocabulary object detection (OVD) addresses this limitation by enabling models to detect novel object categories not explicitly labeled during training. Instead of relying on fixed class labels, OVD methods typically leverage vision-language models trained on image-text pairs to associate regions with arbitrary textual queries [3, 13, 24]. Recent 2D OVD systems such as GLIP [3], Grounding DINO [11], and RegionCLIP [26] demonstrate impressive flexibility and scalability by learning rich semantic representations capable of grounding object categories from natural language.

In this work, we propose a method for open-vocabulary 3D object detection that eliminates the need for manual 3D annotations. Our approach builds upon large-scale open-vocabulary vision-language models trained on 2D images, using their 2D predictions as supervisory signals for 3D detection. This strategy is motivated by two key insights: (i) 2D datasets offer vastly greater object category coverage than their 3D counterparts [5, 9], and (ii) 2D open-vocabulary detectors are significantly more mature and scalable [3, 11].

Without training any new model, our method transfers the generalization ability of 2D vision-language models to

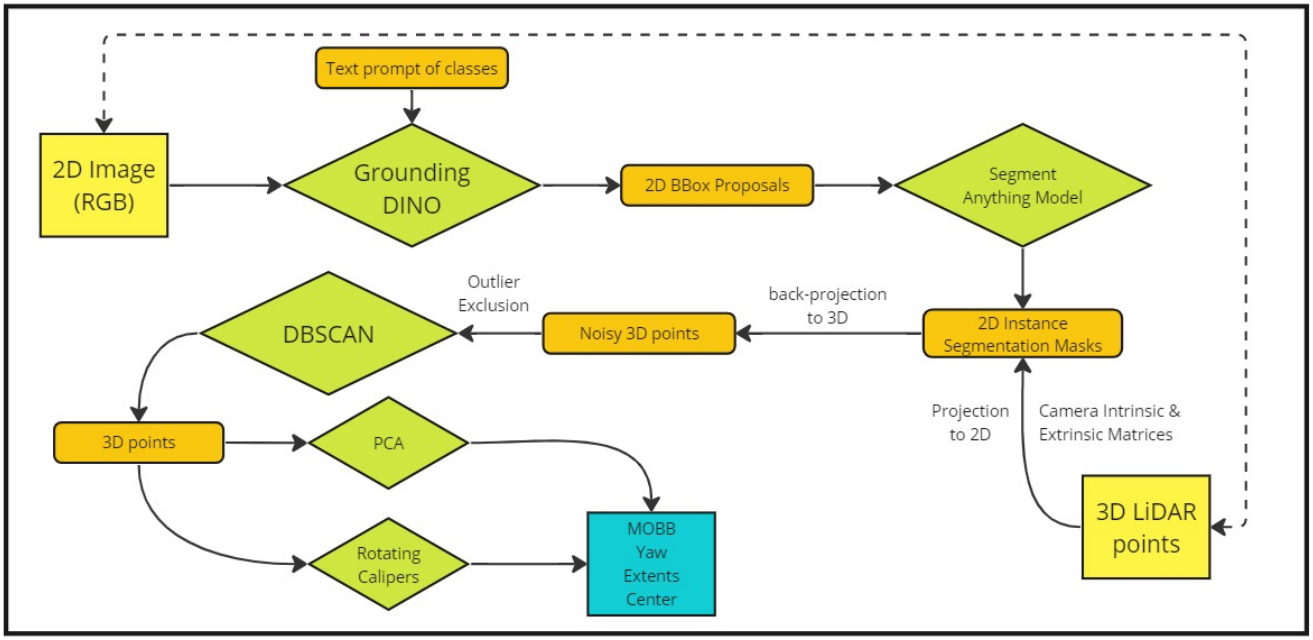

Figure 1. Pipeline of our methodology.

the 3D domain. This enables open-world 3D object detection in previously unseen environments, offering a scalable path toward 3D perception without the burden of dense 3D annotations.

Our key contributions are as follows:

- We propose a novel framework for *open-vocabulary 3D object detection* that requires no manual 3D annotations, leveraging off-the-shelf 2D vision-language models to generate 3D bounding boxes.
- We design a robust 2D-to-3D projection pipeline that incorporates SAM-based segmentation [6], self-supervised noise filtering via DBSCAN, and classical geometric methods to infer 3D bounding box parameters.
- We explore the feasibility of monocular RGB-only 3D detection by introducing a pseudo-LiDAR pipeline using zero-shot depth estimation [15].
- We propose a novel benchmark titled **Pseudo-nuScenes**, to evaluate performance under adverse weather conditions and absence of LiDAR data.
- We conduct comprehensive ablations comparing inflation strategies and demonstrate that our method can produce competitive open-vocabulary 3D predictions under realistic constraints, without any training or dataset-specific priors.

2. Related Work

Open-Vocabulary 2D Object Detection. Open-vocabulary detection (OVD) has gained traction as a way to break free from fixed taxonomies in object detection. Methods like GLIP [3], Grounding DINO [11], and RegionCLIP [26] use large-scale image-text pretraining to associate object regions with arbitrary text queries. These models have demonstrated remarkable zero-shot generalization, making them a powerful tool for object discovery. Our method builds on this foundation using 2D open-vocabulary detectors as the starting point for 3D localization, but differs by targeting 3D reasoning without training or fine-tuning any model.

3D Object Detection and Dataset Limitations. Conventional 3D object detection methods rely on annotated LiDAR or RGB-D point clouds [7, 16, 19]. These models are constrained by small taxonomies (e.g., 10–23 classes in nuScenes [2], Waymo [19], or KITTI [4]), and require costly annotation pipelines. Our approach avoids these limitations by leveraging 2D detections to supervise 3D localization, without requiring any 3D-labeled data.

Open-Vocabulary 3D Detection. A growing body of work explores bringing open-vocabulary capabilities to 3D. Lu *et al.* [14] propose a training-based pipeline that uses 2D open-vocabulary models to generate pseudo-labels and then trains a 3D detector using cross-modal contrastive learning between point clouds, text, and images. CLIP2Scene [12] extends CLIP to 3D by transferring knowledge via 2D-to-3D alignment for indoor scenes. CLIP-FO3D [8] proposes a frozen 3D backbone and aligns 3D proposals with CLIP embeddings for open-set 3D detection. FSD [20] incorporates

CLIP into 3D detectors using a learned 3D-text similarity metric.

Compared to these methods, our pipeline is training-free, does not require any 3D proposals or feature extractors, and operates purely via geometric projection from 2D outputs. Most of the above methods require labeled training data in at least one modality or fine-tuning of encoders, whereas our system is entirely inference-time and model-agnostic.

Zero-Shot 3D Detection via Geometry. Wilson *et al.* [21] (3D for Free) use 2D segmentations and HD maps to generate 3D pseudo-annotations, which are then used to train a 3D detector. However, their method depends on dataset-specific priors and closed-set taxonomies, and is not designed for open-world settings. In contrast, our method uses open-vocabulary 2D outputs, requires no training or HD maps, and generalizes to novel object categories via natural language.

Vision-Language Models for 3D Understanding. Several works use CLIP-like models for 3D representation learning. PointCLIP [25] adapts CLIP to classify point clouds via projection. ULIP [23] aligns point cloud, image, and language embeddings into a unified space via contrastive learning. While related in spirit, these methods primarily focus on classification and require extensive multimodal pretraining. In contrast, our approach treats pretrained 2D vision-language models as fixed oracles and focuses on fully automated 3D detection.

Summary. In summary, our method is the first to demonstrate fully training-free, open-vocabulary 3D object detection using only 2D foundation models and classical geometric reasoning. It avoids the need for dataset-specific priors, annotations, or learned 3D feature extractors, enabling scalable 3D perception from purely 2D supervision.

3. Methodology

We propose a zero-shot pipeline for open-vocabulary 3D object detection that leverages 2D vision-language models and 3D geometric reasoning, while requiring no human-labeled 3D annotations. Our approach is summarized in Figure 1 and consists of five key stages: 2D open-vocabulary detection, instance segmentation, 2D-to-3D back-projection, geometric box inflation, and optional pseudo-depth substitution.

3.1. Stage 1: 2D Open-Vocabulary Detection

Given a text prompt specifying target object classes, we first run a vision-language detector such as GroundingDINO [10] on the input RGB image. The model returns a set of text-conditioned 2D bounding boxes with associated class labels and confidence scores. These detections serve as the initial priors for subsequent 3D reasoning.

3.2. Stage 2: Instance Segmentation via SAM

To improve spatial localization and minimize noise in 3D point selection, we apply Segment Anything (SAM) [6] to generate high-quality instance masks. We condition SAM on the bounding boxes predicted by GroundingDINO, producing fine-grained object-level segmentation masks.

3.3. Stage 3: LiDAR-to-Image Projection and Back-Projection

Next, we project the LiDAR point cloud into the image plane using the known camera intrinsics and extrinsics. This establishes a pixel-wise mapping between 2D pixels and 3D points. For each segmentation mask, we retrieve the corresponding 3D points by inverting this mapping, effectively *lifting* the 2D instance mask into a 3D point cloud segment. However, this set of points may be noisy due to projection ambiguities and imperfections in the mask.

3.4. Stage 4: 3D Bounding Box Inference

To convert the masked 3D point cloud into a 3D bounding box, we explore several inflation strategies:

- **Medoid-based Centering:** We compute the medoid of the 3D points as a robust estimate of the box center.
- **Rotating Calipers for Orientation and Size:** We adapt the classic Rotating Calipers algorithm to fit a minimum-area oriented bounding box to the point cloud in the ground plane. The box's height is derived from the vertical extent of the points.

We compare the effectiveness of various combinations of these strategies (e.g., medoid center with shape priors vs. full Calipers inflation) in Section 5.

To reduce the impact of outliers, we apply DBSCAN clustering on the back-projected 3D points and retain only the densest cluster, yielding cleaner box proposals.

3.5. Stage 5: Pseudo-Depth Variant (RGB-Only)

In scenarios where LiDAR is unavailable, we substitute real depth data with pseudo-LiDAR generated from RGB images using UniDepth [15], a zero-shot monocular depth estimator. We back-project the RGB pixels into 3D using the estimated depth map and intrinsic parameters. The rest of the pipeline remains identical. We evaluate this variant on our augmented Pseudo-NuScenes dataset, described in Section 4.

3.6. Label Assignment

For each predicted 3D box, we transfer the class label and confidence score from the original 2D detection. No additional training or fine-tuning is performed, and our pipeline

remains entirely model-agnostic with respect to the downstream 3D task.

4. Pseudo-nuScenes: Fog-Augmented RGB-D Benchmark

Recent research in 3D object detection has highlighted the limitations of relying solely on LiDAR data: high cost, limited accessibility, and sensitivity to weather. Moreover, most existing datasets fail to account for real-world degradations such as fog or occlusion. To study the effectiveness of open-vocabulary 3D object detection in such conditions, we introduce a novel benchmark: **Pseudo-nuScenes**.

4.1. Dataset Construction

Pseudo-nuScenes is derived from the official mini split of the nuScenes dataset [1], which contains over 1000 annotated frames from synchronized RGB and LiDAR sensors. We discard all LiDAR signals and instead generate a pseudo-3D structure from monocular RGB using state-of-the-art zero-shot metric depth estimation.

Monocular Depth via UniDepth. For each RGB image, we use UniDepth [15], a universal monocular depth model, to infer metric depth maps. These depth maps are projected into 3D using known intrinsic parameters from the nuScenes calibration data. The resulting pseudo-point clouds mimic LiDAR scans and can be processed by our 3D inflation pipeline as drop-in replacements.

Depth Generation. For each RGB image from the six nuScenes cameras, we run UniDepth to obtain per-pixel metric depth maps. These maps are projected into 3D using the known camera intrinsics, producing dense pseudo-point clouds. The resulting 3D structure allows us to simulate LiDAR-like observations for each frame without requiring any additional sensor data.

Fog Augmentation. To simulate real-world visibility degradation, we apply a physics-inspired fog model to the RGB images. The model attenuates pixel intensities based on scene depth using the standard exponential transmittance function:

$$I_{\text{fog}}(x) = I(x) \cdot t(x) + A \cdot (1 - t(x)), \quad t(x) = e^{-\beta d(x)}$$

where $I(x)$ is the original pixel intensity, $d(x)$ is the depth at pixel x, β is the fog density, and A is the ambient atmospheric light (set to white). This simulates realistic fog effects where distant regions become low contrast and desaturated.

Setup. We process all frames from the mini split of nuScenes, resulting in a fully RGB-D dataset that can be used to test open-vocabulary 3D detection pipelines under more challenging conditions. Our dataset is particularly useful for evaluating methods in domains where active sensors (e.g., LiDAR) are unavailable, unreliable, or prohibitively expensive.

5. Experiments

5.1. Datasets

nuScenes. We evaluate our method on the mini-split of the nuScenes dataset [1], a large-scale multimodal benchmark for autonomous driving. The dataset includes synchronized RGB images from six cameras, LiDAR scans, and 3D bounding box annotations across 23 object classes. For our experiments, we use only the two validation sequences (10 scenes total) due to our model-free setup and compute constraints.

Pseudo-nuScenes (Ours). We introduce a fog-augmented, RGB-D variant of nuScenes, named *Pseudo-nuScenes*, to study the robustness of our method in RGB-only settings. We discard all real LiDAR and generate pseudo-LiDAR from RGB images using UniDepth [15], a zero-shot monocular metric depth estimator. We additionally apply synthetic fog augmentations using a depth-aware haze simulation to mimic adverse weather. Our dataset allows us to evaluate open-vocabulary detection under degraded, real-world visual conditions.

5.2. Evaluation Metrics

We follow the nuScenes detection benchmark protocol [2]. Our primary metric is:

- **Mean Average Precision (mAP):** We follow the official definition, where a match is based on the 2D Euclidean distance of ground-plane box centers, averaged over thresholds $\{0.5, 1, 2, 4\}$ meters.

We also report the nuScenes suite of true positive (TP) quality metrics:

- **mATE**: Mean Average Translation Error (m).
- **mASE**: Mean Average Scaling Error.
- **mAOE**: Mean Average Orientation Error (radians).
- **mAVE**: Mean Average Velocity Error (m/s).
- **mAAE**: Mean Average Attribute Error.
- **mAR**: Mean Average Recall.
- **NDS**: NuScenes Detection Score, combining all the above via a weighted average.

5.3. Baselines and Variants

We compare our method with:

- **3D For Free** [21]: A prior method that inflates 2D segmentations to 3D using HD maps and hand-crafted priors.
- **Ours (Medoid + Shape Priors)**: Uses LiDAR medoid as box center and fixed class-specific shape priors.
- **Ours (Rotating Calipers)**: Uses 3D Rotating Calipers for box orientation and size; we ablate center strategies.
- **Ours (Pseudo-Depth)**: Fully RGB-only pipeline using UniDepth + fog augmentation from Pseudo-nuScenes.

Table 1. Comparison of inflation strategies on **nuScenes**. Metrics are reported on the 5 most common classes.

Method	DBSCAN	mAP	mATE	mASE	mAOE	mAVE	mAAE	mAR	NDS
Medoid + Lane geometry	Yes	**29.94%**	0.938	0.700	**1.045**	1.560	0.982	41.78%	**18.77%**
+ shape priors	No	29.42%	0.948	0.700	1.045	1.558	0.982	40.24%	18.41%
Rot. Calipers (center,	Yes	**21.94%**	0.956	0.879	**1.155**	1.566	0.980	36.74%	**12.82%**
orientation, shape)	No	1.30%	1.029	0.977	1.144	1.151	0.990	6.76%	0.99%
Medoid + Rotating Calipers for orientation, shape	Yes	**29.30%**	0.949	0.897	**1.155**	1.552	0.981	40.10%	**16.38%**
3D For Free w/ HD maps	-	37.40%	0.41	0.31	0.90				
3D For Free w/ Rot. Calipers	-	34.31%	0.54	0.33	1.35				

Table 2. Performance on Pseudo-nuScenes (fog + UniDepth).

Method	**mAP**	**mATE**	**mASE**	**mAOE**	**mAVE**	**mAAE**	**NDS**
Medoid + Shape Priors + Lane Geometry	16.14	1.053	0.703	1.039	1.545	0.981	11.23
Rot. Calipers (center, orientation, shape)	12.21	1.060	0.890	1.148	1.484	0.980	7.40

6. Results and Discussion

6.1. Impact of Inflation Strategy

Table 1 presents results on the nuScenes mini validation set. We find that using the medoid as the center, along with handcrafted lane geometry and shape priors, achieves an mAP of 29.94%, closely matching the performance of the prior 3D For Free baseline (34.31%). Our method, in contrast, uses no dataset-specific priors or HD maps, and supports open-vocabulary queries.

We observe that the mASE scaling error increases by 0.2 between the shape priors baseline and both versions of rotating calipers. This increase is expected because the shape priors, by design, are hand-crafted features intended to generate more appropriately sized anchor boxes. Interestingly, our method achieves equivalent translation performance as the shape priors baseline. Recall, attribute error, and velocity error remain similar across the various methods as the inflation method has little consequence on these factors.

However, consider the orientation error (mAOE). Our method featuring the 3D Rotating Calipers strategy yields improved orientation estimates, with a lower mAOE (1.045 vs. 1.144), **surpassing the baseline** inflation methods [21]!

However, this method suffers in mAP and recall unless combined with medoid-based centering. This highlights the importance of combining robust shape estimation with reliable center inference.

6.2. Noise Suppression via Clustering

To evaluate the effect of noise in projected 3D points, we apply DBSCAN for outlier removal. As shown in Figure 2, the noisy predictions result in severely degraded box quality. With DBSCAN, the mAP improves from 1.30% to 21.94%, affirming the value of density-based filtering.

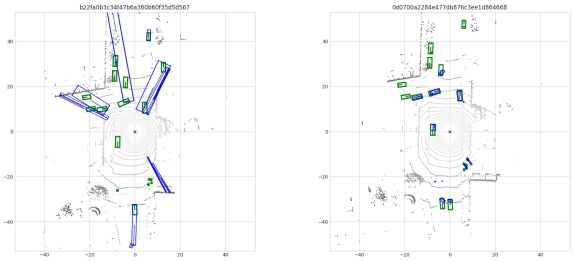

Figure 2. Effect of DBSCAN-based noise filtering on BEV maps. **Left**: Without filtering. **Right**: After DBSCAN. Blue: prediction, Green: ground truth.

6.3. Performance in RGB-Only Settings

In Table 2, we show results on our Pseudo-nuScenes dataset. As expected, replacing LiDAR with pseudo-depth leads to a drop in accuracy, especially in mAR and orientation. Nevertheless, our method achieves non-trivial mAP values (12–16%), suggesting potential for fully vision-only setups in constrained environments.

However, the translation error (mATE) remains roughly similar between the pseudo-depth predictions and the actual depth predictions. Likewise, scaling error (mASE) is equivalent, implying that the predicted boxes have equivalent performance in terms of box scale and positioning. However,

there is significantly less recall using pseudo-depth. The predicted depth labels appear to lead to noisier/more confusing model predictions than the real depth labels, leading to lower recall.

6.4. Qualitative Results

Figure 3 shows BEV visualizations. Our method captures major structures accurately in LiDAR-rich setups. In RGB-only settings, the predictions become noisier but still capture coarse layout.

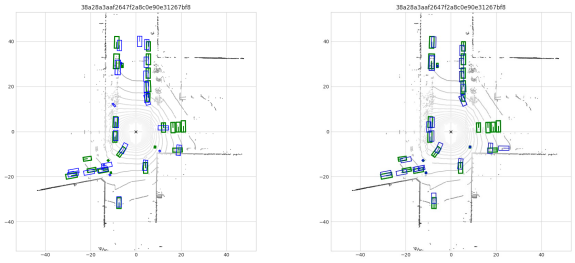

Figure 3. Qualitative results on nuScenes (left) and Pseudo-nuScenes (right). Blue: prediction, Green: ground truth.

7. Conclusion

We present a new paradigm for open-vocabulary 3D object detection that mitigates the need for human-labeled 3D annotations. The core principle underlying our work is that 2D vision-language models trained on web-scale, richly diverse datasets possess semantic understanding and object localization capabilities that far exceed what current 3D datasets and detectors can offer. By harnessing these models as the supervisory backbone, we lift 2D detections to 3D space using classical geometric reasoning and minimal prior assumptions. Additionally, using our method leverages the larger taxonomies and maturity of 2D datasets.

Our pipeline remains entirely training-free, modular, and open-vocabulary, supporting flexible object queries and generalization to novel categories. To evaluate the real-world viability of this approach, we also introduce **Pseudo-nuScenes**, a fog-augmented, RGB-only benchmark derived from nuScenes, which simulates degraded visibility and absence of depth sensors.

Extensive experiments show that:

- 2D detectors can serve as surprisingly strong priors for 3D object discovery,
- Our inflation strategy using segmentation masks, DBSCAN filtering, and Rotating Calipers achieves competitive localization performance,
- Monocular pseudo-depth provides a feasible (albeit noisy) alternative to LiDAR in open-world 3D settings.

By transferring the strengths of 2D detection to the 3D domain, our work takes a step toward scalable, annotation-free, open-vocabulary 3D perception. We hope this direction inspires future research on bridging the modality gap through 2D-centric supervision, especially as 2D foundation models continue to improve.

Acknowledgements

We would like to express our sincere gratitude to the Vision Lab at the Infosys Centre for AI, Indraprastha Institute of Information Technology, Delhi, for their generous support in providing us an NVIDIA A100 GPU for conducting our experiments. We also extend our thanks to Dr. Saket Anand for his mentorship as our bachelor's thesis advisor. His support was instrumental in helping us develop the skills necessary to execute this independent research.

References

[1] 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. 4

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2, 4

[3] Xiao Du, Hong Zhang, Zhen Li, Xiangyang Lin, et al. Glip: Grounded language-image pre-training. In *CVPR*, 2022. 1, 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2

[5] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 3

[7] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1, 2

[8] Zhengxiong Li, Qian Ye, Tianrui Wang, et al. Clip-fo3d: Free open-vocabulary 3d object detection. In *ICCV*, 2023. 2

[9] 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*, pages 740–755, Cham, 2014. Springer International Publishing. 1

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

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[12] Yujing Liu, Wenhao Xu, et al. Clip2scene: Scene-level 3d open-world understanding via vision-language foundation models. In *CVPR*, 2023. 2

[13] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1190–1199, 2023. 2

[15] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[16] Charles R Qi, Wei Liu, Chen Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 1, 2

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[19] Pei Sun, Henrik Kretzschmar, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 2

[20] Yuchen Wang, Qi Zhou, Jianwei Yang, et al. Fsd: Few-shot object detection in 3d scenes. In *CVPR*, 2023. 2

[21] Benjamin Wilson, Zsolt Kira, and James Hays. 3d for free: Crossmodal transfer learning using hd maps. *arXiv preprint arXiv:2008.10592*, 2020. 3, 4, 5

[22] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes*, 2018. 1

[23] Yujing Xue, Yue Wang, Xingyu Liu, et al. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. In *ECCV*, 2022. 3

[24] Alireza Zareian, Kevin D Wang, Roozbeh Mottaghi, Ali Farhadi, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 1

[25] Yujing Zhang, Enze Xie, Jiwen Dai, and Zhaoxiang Yu. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 3

[26] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 2