Open-Vocabulary 3D Object Detection Without 3D Human Annotations

Atharv Goel* Mehar Khurana* Prakhar Gupta
IIIT Delhi IIIT Delhi IIIT Delhi
{atharv21027, mehar21541, prakhar21550}@iiitd.ac.in

Department of Computer Science & Engineering, IIIT Delhi

Abstract

3D object detection is a crucial task in computer vision. However, modern 3D datasets for detection often lack the diversity of 2D datasets, hindering the generalization of 3D detectors across objects and scenarios. To address this, we work on leveraging the robustness and size of 2D datasets for 3D tasks and implementing open-vocabulary detection to detect novel classes from the dataset. Our work presents an algorithm for open-world 3D object detection without requiring human annotations, capitalizing on the recent advancements in vision-language foundation models trained on 2D datasets. By extending the capabilities of 2D object detection models to the domain of 3D object detection, we aim to facilitate the detection of novel instances in a 3D environment without the need for manual 3D annotations, thereby bridging the gap between 2D and 3D perception. We make our code and resources available at https://github.com/Zynade/open-world-3D-det.

1. Introduction

Object detection is a fundamental task in computer vision, essential for identifying and localizing objects within images or scenes. It underpins various applications such as autonomous vehicles, surveillance systems, augmented reality, and robotics. Traditionally, object detection methods have focused on analyzing objects in 2D images. However, with technological advancements, there's a growing demand for more sophisticated perception capabilities, leading to the development of 3D object detection techniques. Unlike their 2D counterparts, 3D object detection methods aim to perceive objects in three-dimensional space, capturing their appearance, spatial layout, and geometry.

The limited class taxonomies in modern 3D datasets for detection pose a challenge as they lack the diversity and richness of 2D detection datasets. This disparity restricts the ability of 3D detectors to generalize across a wide range of objects and scenarios, leading to sub-optimal performances in real-world applications. Hence, it becomes imperative to explore the potential methods of exploiting the robustness and size of the 2D datasets for 3D tasks.

Open vocabulary detection aims to detect novel classes that are not labelled in the training dataset. Traditional detection models are trained on a fixed set of object classes, restricting their ability to generalize to novel classes. In contrast, open-vocabulary detection methods leverage techniques such as image-text pairs or vision-language embedding to learn a general representation of what objects look like, allowing it to detect objects based on arbitrary text descriptions. This approach enables the model to perform detection tasks without prior knowledge of the specific object classes, making it particularly useful for scenarios where the object classes may be diverse or evolving.

This work aims to develop an algorithm for open-world 3D object detection without requiring human annotators. The aim is to capitalize on recent advancements in large-scale open-vocabulary vision-language foundation models. These models, primarily trained on 2D datasets, can detect objects and output 2D bounding boxes.

The rationale behind this approach is twofold: firstly, to exploit the extensive variety of object categories available in 2D datasets, and secondly, to leverage the relative maturity of 2D object detectors as compared to 3D detectors and datasets.

Our work seeks to extend the capabilities of 2D object detection models to the domain of 3D object detection, thereby facilitating the detection of novel instances in a 3D environment without the need for manual 3D annotations.

Further, we achieve this goal without training any model.

2. Related Work

Wilson *et al.* [6] fuse 2D segmentation masks from a regular closed-set instance segmentation model and back-project these points to 3D world space. Their method requires

*These authors contributed equally to this work

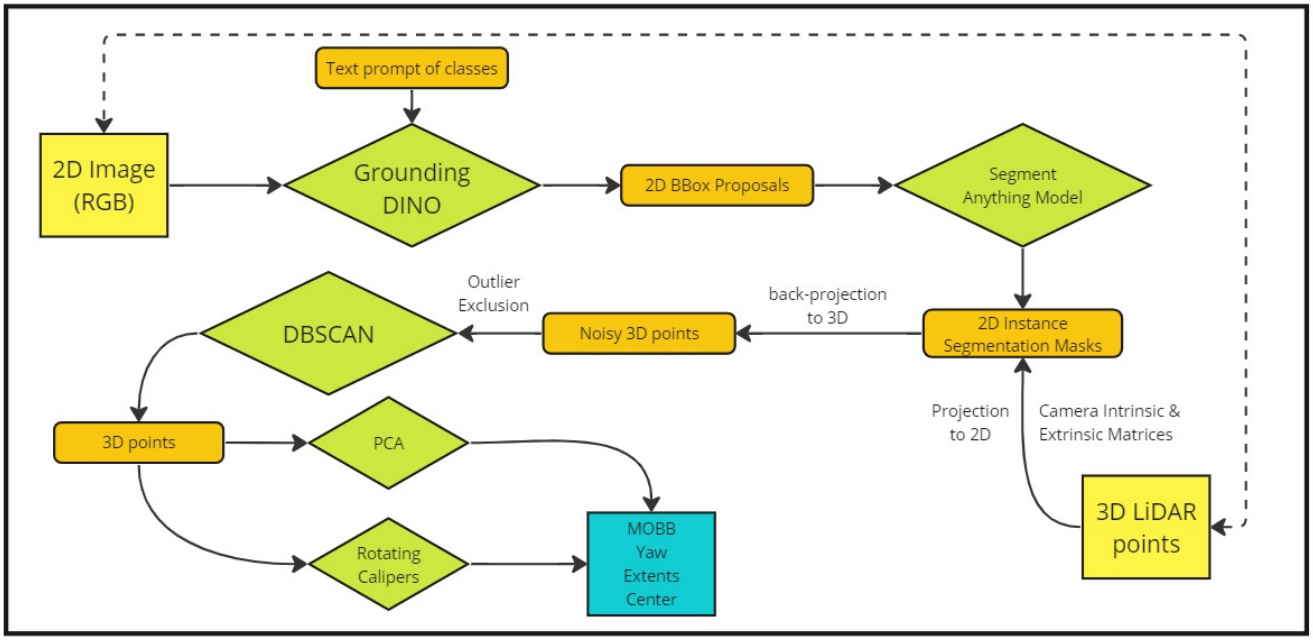


Figure 1. Pipeline of our methodology.

RGB-LiDAR correspondence pairs. Using this mining process, they estimate the 3D bounding boxes by constraining the 2D-to-3D mapping with pre-defined class priors and HD maps from the dataset. They then train a detector on these predictions as pseudo-annotations.

However, their method has three major limitations:

1. Their method heavily relies on the pre-defined class priors and HD maps, which makes it not generalizable to different datasets.

2. The method fails to work in the case of a lack of RGB-LiDAR correspondence. LiDAR is expensive and it is not always feasible to obtain the required LiDAR data.

3. Their model is restricted to a closed taxonomy. This is a harsh restriction in the real world with varying classes of objects that may not be specified in a closed vocabulary setting.

Lu *et al.* [4] proposes an open-world point-cloud detector that learns localization from pseudo-ground truths from the 2D foundation model. For learning classification, it proposes a triplet cross-modal contrastive learning mechanism to incorporate information from all three modalities: RGB, LiDAR, and text.

3. Methodology

Our method is strongly inspired from the work of Wilson *et al.* [6]. We build over their work, seeking to solve the three limitations of their work as described in Section 2.

Our method is visualized in a neat flowchart in Figure 1.

Our method works best with RGB-LiDAR correspondences as input. The RGB image is passed through a foundation model like GroundingDINO [3] to generate initial 2D boxes, with a text prompt containing the desired classes to be detected. Next, we use SAM [2] conditioned on these 2D proposals to generate instance segmentation masks for each object. We then project the 3D LiDAR points into pixel space using the camera's extrinsic and intrinsic parameters, storing this projection map for future use.

In the next step, we back-project the 2D masks to 3D points using the LiDAR-pixel space mapping stored in the previous step. However, this mining process is noisy, so we perform DBSCAN clustering to filter out outlier points. To identify the right box orientation, we use the Rotating Calipers method, a classical technique from computational geometry. We apply Rotating Calipers in 3D to refine the box's yaw, pitch, and roll.

The box center could be computed directly from the Rotating Calipers algorithm, or we could compute the medoid of the 3D points. We conduct experiments to analyze the effectiveness of either approach. For classification, we assign the category label and confidence score predicted by GroundingDINO.

In the absence of LiDAR data, we use UniDepth [5], a zero-shot monocular metric depth estimator, to estimate the depth map for the RGB image. We project these image points to 3D space. This data serves as a pseudo-LiDAR data source, which can be fed into the algorithm like regular LiDAR data.

Table 1. Comparing various inflation methods on **nuScenes**. We report the mAP for our method on the 5 most common classes in nuScenes.

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%
Rotating Calipers for 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 [6] w/ HD maps	-	37.40%	0.41	0.31	0.90				
3D For Free [6] w/ Rot. Calipers	-	34.31%	0.54	0.33	1.35				

4. Experiments

4.1. Experimental Setup

Datasets. To evaluate our method, we use the mini split of the nuScenes dataset [1]. NuScenes is a standard benchmark dataset for autonomous driving, containing over 1000 scenes from Boston and Singapore, containing 20-second sequences. The data is collected from 6 cameras situated on a vehicle and a LiDAR sensor on the top. The dataset features 23 object classes with 3D bounding boxes annotated on the images from each camera, neatly synchronized to samples within the sequences. However, due to computing and time restrictions, we use the mini split of NuScenes, which has 10 scenes – 8 scenes for training and 2 for validation. Since we do not perform training, we use only the two validation sequences.

4.1.1 Pseudo NuScenes

We also use an augmented version of NuScenes for our RGB-D experiments. We replace all LiDAR data with pseudo-LiDAR data generated purely from RGB images of the 6 camera modalities and add fog augmentations to the RGB images. We call this augmented dataset ***Pseudo NuScenes***. More information in Section 3.

4.1.2 Evaluation Metrics.

- **Mean Average Precision (mAP)**: This is classic mAP, but instead of IoU for localization, a match is defined by considering the 2D center distance on the ground plane. Specifically, we match predictions with the smallest center distance up to a certain threshold. We calculate average precision (AP) for a given match threshold by integrating the precision-recall curve for recalls and precisions > 0.1. We finally average over thresholds of 0.5, 1, 2, 4 meters and compute the mean across classes.

NuScenes defines the following metrics for assessing the quality of True Positive (TP) predictions. The following information is sourced from their webpage [1].

- **Mean Average Translation Error (mATE)**: 2D Euclidean center distance (in meters).
- **Mean Average Scaling Error (mASE)**: computed as $1 - IoU$ after aligning centers and orientation.
- **Mean Average Orientation Error (mAOE)**: Smallest yaw angle difference between predictions and ground truths in radians.
- **Mean Average Velocity Error (mAVE)**: Absolute velocity error in meters/second.
- **Mean Average Attribute Error (mAAE)**: computed as $1 - acc$, where acc is the attribute classification accuracy.
- **Mean Average Recall (mAR)**: The standard mAR metric – the maximum recall given a fixed number of detections per image, averaged over categories and IoUs.

  NuScenes neatly incorporates all of the information from these metrics into a single concise metric called the NuScenes Detection Score.
- **NuScenes Detection Score (NDS)**: NDS is computed using a weighted average of the TP metrics as follows:

$$\textbf{NDS} = \frac{(5 \cdot \text{mAP} + \text{mATE} + \text{mASE} + \text{mAOE} + \text{mAVE} + \text{mAAE})}{10}$$

Baselines. We compare our method with the inflation method of [6]. We compare against their inflation using HD maps and box priors and inflation using Rotating Calipers. We also generate our own baseline using 3D Rotating Calipers and box priors.

4.2. Results

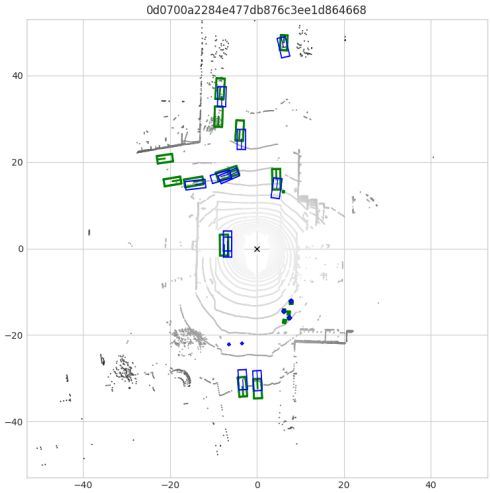
We conduct ablation studies based on the inflation method.

1. Lane geometry and shape priors
2. Rotation Calipers for box orientation and size; medoid for center
3. Rotation Calipers for box orientation, size, center

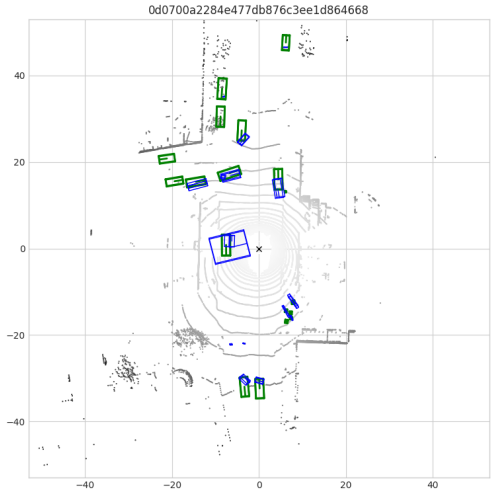
Our results on NuScenes are summarized in Table 1.

Table 2. Comparing various inflation methods on our **augmented nuScenes with Pseudo-Depth**. We report the mAP for our method on the 5 most common classes in nuScenes.

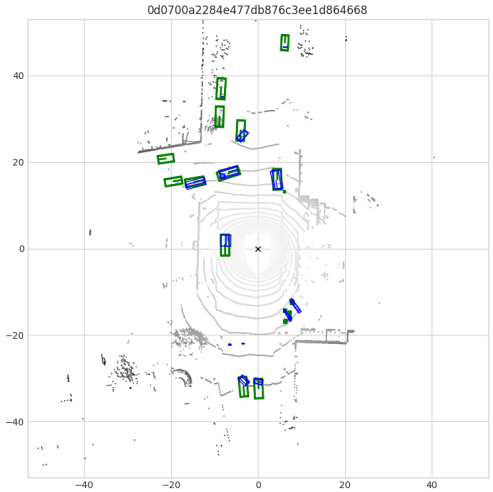
Method	DBSCAN	mAP	mATE	mASE	mAOE	mAVE	mAAE	mAR	NDS
Medoid + Lane geometry + shape priors	Yes	16.14%	1.053	0.703	1.039	1.545	0.981	29.04%	11.23%
Rotating Calipers for center, orientation, shape	Yes	12.21%	1.060	0.890	1.148	1.484	0.980	26.82%	7.40%



(a) Medoid + Lane geometry + shape priors

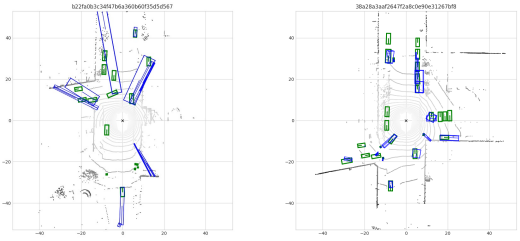


(b) Medoid + Rotating Calipers for orientation, shape



(c) Rotating Calipers for center, orientation, shape

Figure 2. Bird's Eye View (BEV) HD Maps computed using our method on NuScenes. Blue - predictions; Green - ground truths.



(a) No noise removal (b) With noise removal

Figure 3. BEV HD Maps of our method on NuScenes: 3D Rotating Calipers for orientation, size, box center. **Blue - predictions; Green - ground truths.**

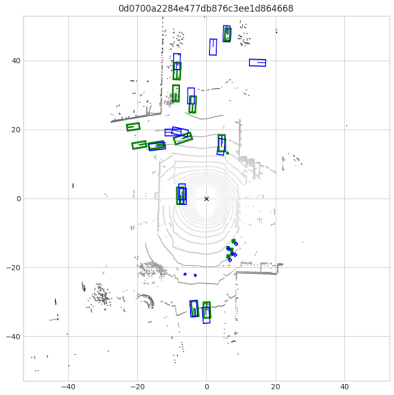


Figure 4. Bird's Eye View (BEV) HD Map computed on **Pseudo-NuScenes**, our augmented version of NuScenes featuring Pseudo-Depth labels. This figure visualizes the lane geometry + shape priors baseline, featuring our method for estimating box center using cloud medoid. **Blue - predictions; Green - ground truths.**

4.2.1 Inflation method

Table 1 shows that our method utilizing 3D rotating calipers for orientation and shape and point cloud medoid for box center achieves the same performance as the hand-crafted closed-vocabulary baseline of Lane geometry + shape priors. This mAP is about 29.4%, which has competitive value with the method of Wilson *et al.*[6] achieving 34.31%.

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). This is the smallest yaw angle difference between a prediction and ground truth. Our method featuring 3D rotating calipers achieves very impressive results, **surpassing the baseline** of 3D For Free's inflation method!

Qualitative Analysis. We visualize the Bird's Eye View (BEV) HD maps visualized in Figure 2. We observe much better matches between GT and prediction using medoid + lane geometry + shape priors than we do for rotating calipers. Once again, this is expected due to the design of the hand-crafted baseline. Moreover, the rotating calipers appears to have many more samples without a matched prediction, i.e., False Negatives. However, when we use point cloud medoid as box center instead of the rotating calipers method, we observe much tighter matches, which agrees with our quantitative observation from Table 2.

4.2.2 Noise suppression

We experiment with removing outliers from the inflated 3D points using the density-based clustering algorithm DBSCAN.

The inflated boxes generated from the pipeline may be noisy due to the ill-posed 2D to 3D projection. We perform outlier exclusion using clustering to verify if removing such outlier points improves performance.

Table 1 shows that when we use 3D rotating calipers to estimate orientation, shape, and box center, the method achieves a mere 1.30% mAP and only 6.76% mAR. Obviously, this is unusably poor. This observation indicates that the inflated boxes are very poor. However, when we introduce DBSCAN clustering, the mAP shoots up to nearly **22%**. This observation indicates that the inflated boxes have an extreme amount of noise in them, and the density-based outlier removal is extremely effective in de-noising these predictions.

Qualitative Analysis. We observe the BEV HD maps in Figure 3. The noisy maps are extremely poor quality, and seemingly random.

4.2.3 Pseudo-NuScenes

We evaluate our method for purely RGB images using our **Pseudo-NuScenes** dataset.

Table 2 shows the results for Pseudo-NuScenes. We observe a significant decrease in performance from real LiDAR data, which is expected due to the noisy process of monocular metric depth estimation. We observe 16% mAP using lane geometry and shape priors versus 12% mAP using 3D rotating calipers for center, orientation, and shape. 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.

5. Conclusion

Atharv. Here's what I conclude from my analysis. Our method observes competitive performance with competing baselines, however, it falls short too many times to be usable in its current state. The method works in some specific conditions, but not very well in most other conditions, rendering it unreliable. For instance, using rotating calipers plus point cloud medoid as the box center achieves the same performance as using lane geometry information, which is a closed-vocabulary baseline. Our method achieves similar performance in this case with the added benefit of allowing an open-vocabulary setting.

We find that using RGB-only images without actual depth information does not work as we hoped it would. The metric depth estimation from monocular images is an ill-posed mapping, rendering it considerably less effective than actual depth information. We require depth information either in the form of LiDAR point clouds or depth maps for such an approach to work.

More generally, we have established that it is possible to achieve open-vocabulary 3D object detection without requiring a human to annotate any 3D data at all. We can simply use well-established off-the-shelf 2D detectors and instance segmentation models to achieve reasonable predictions in 3D. Which, in and of itself, is impressive. There may be use cases where the massive annotation cost of annotating a huge 3D dataset is not worth the performance gain derived from it.

Mehar. With this project, we try to establish a comparison between the State-of-the-art in 2D foundational object detectors and 3D (domain-specific) object detectors. We find that simply inflating predictions from 2D foundational object detectors like GroundingDINO achieves close-to-SOTA numbers with the added benefit of being open-vocabulary. Through our ablative analysis we also find the optimum strategy for said "inflation" – which is to use the medoid of the masked LiDAR points along with lane and shape priors (which are closed-vocabulary). As can be seen in the BEV images, rotating calipers largely fails due to the high possibility of outliers in masked LiDAR points, resulting from depth-discontinuities at the boundary of the instance masks and minor calibration errors between the RGB and LiDAR data. We also find that our method performs poorly when using pseudo-depth and augmented RGB data, highlighting the noise in both the monocular metric depth estimator, UniDepth [5], and the zero-shot detector [3].

Prakhar. In this project, we tried to utilise mature 2D object detectors instead of 3D object detectors for 3D object detection. The experiments conducted highlight the potential of this method and the challenges of using synthetic depth data derived from RGB images for object detection. The 3D rotating calipers method achieves competing results compared to traditional LiDAR-based approaches using lane geometry and shape priors, with a general improvement in bounding box estimation with noise removal using DBSCAN, which leads to a significant improvement in detection metrics. This suggests that refining pre-processing and noise removal in pseudo-depth data has merit and can be useful for the improvement of performance.

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 computational resources. Their assistance in supplying us with an NVIDIA A100 GPU significantly helped us conduct our experiments efficiently. We also extend our thanks to Prof. Saket Anand for his guidance and mentorship as our bachelor's thesis advisor. His support was instrumental in helping us develop the skills necessary to undertake this independent study.

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. 3

[2] 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] 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. 2, 6

[4] 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

[5] 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, 6

[6] Benjamin Wilson, Zsolt Kira, and James Hays. 3d for free: Crossmodal transfer learning using hd maps. *arXiv preprint arXiv:2008.10592*, 2020. 1, 2, 3, 5