Supplementary Materials: Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving

Mahyar Najibi* Jingwei Ji* Yin Zhou† Charles R. Qi Xinchen Yan
Scott Ettinger Dragomir Anguelov
Waymo LLC

1. Implementation Details

This section provides the implementation details for the main two components of our approach namely multi-modal auto labeling, and the open-vocabulary 3D object detector.

1.1. Unsupervised Multi-modal Auto Labeling

In our experiments, we use VEGETATION, ROAD, STREET, SKY, TREE, BUILDING, HOUSE, SKYSCRAPER, WALL, FENCE and SIDEWALK as text queries for defining background categories, C^{bg}, which are excluded from auto labeling. We also set the cosine similarities threshold ϵ^{bg} to be 0.02. For the experiments in Section 4.2, and 4.3.1 of the main paper which consider moving-only objects, we set a scene flow threshold of $\epsilon^{sf} = 1m/s$ (the same as [4]). For bounding box proposals, we follow Najibi *et al.* [4] and set neighborhood threshold to be 1.0m in DBSCAN. Without knowing the semantics of objects, it is challenging to define the headings of all objects. For moving objects, we align their headings with the object moving direction. For static objects, we choose their headings such that they have an acute angle with the heading of the autonomous driving vehicle.

1.2. Open-vocabulary 3D Object Detection

Regarding the vision-language model, in this paper we use the pre-trained OpenSeg model [1] coupled with the BERT-Large text encoder in Jia *et al.* [3] without further fine-tuning on any 2D or 3D autononmous driving datasets.

For the knowledge distillation, as discussed in Section 3.3.2 of the main paper, we directly distill the final 640 dimensional features of the OpenSeg model. However, for memory and compute efficiency during training, we first reduce the dimensionality of the features to 64 using an incremental PCA fitted to the whole unsupervised training dataset. To evaluate the open-vocabulary detector on the Waymo Open Dataset, we choose the vehicle and VRU as categories of interest, for which the dataset has groundtruth. More specifically, we use CAR, VEHICLE, PARKED VEHICLE, SEDAN, TRUCK, BUS, VAN, MINIVAN, SCHOOL BUS, PICKUP TRUCK, AMBULANCE, FIRE TRUCK to query for the vehicle category and CYCLIST, HUMAN, PERSON, PEDESTRIAN, BICYCLE to query for the VRU category. We found that removing queries from this set will lead to dropped mAPs. For the 3D detection experiments, we use the same two-frame anchor-based PointPillars backbone as previous work [4] for fair comparisons. We also use the same set of detection losses to train a class-agnostic 3D bounding box regression branch and an objectness score branch, and supplement them with the new distillation introduced in Section 3.3.2 of the main paper. We train models on 64 TPUs, with a batch size of 2 per accelerator. We use a cosine decay learning rate schedule and an initial learning rate of 0.003 and train the models for a total of 43K iterations.

2. Additional Qualitative Results

In the paper, we presented qualitative results demonstrating that UP-VL can detect open-set objects using text queries at inference (see Figure 1 and 4 of the main paper). Additionally, we included a quantitative comparison with the previous state-of-the-art, MI-UP [4], in Table 1 of the main paper. Here, we present qualitative comparison between our UP-VL detector (trained with distillation) and MI-UP [4] detector in Figure S1. The top row shows our UP-VL class-aware predictions where the blue and red boxes represent the vehicle and VRU detections respectively. On the bottom, we are showing the class-agnostic predictions of the MI-UP model as green boxes. Comparing column (a), first we can see that unlike MI-UP which is unable to predict semantics, our UP-VL approach can reliably distinguish between objects of vehicle and VRU categories. Moreover, UP-VL can detect many of the objects which were completely missed or grouped together by MI-UP. In column (b), we also mark static objects in the bottom row. Comparing this column highlights another advantage of our approach. While MI-UP is limited to detect-

*Equal contribution
†Corresponding author

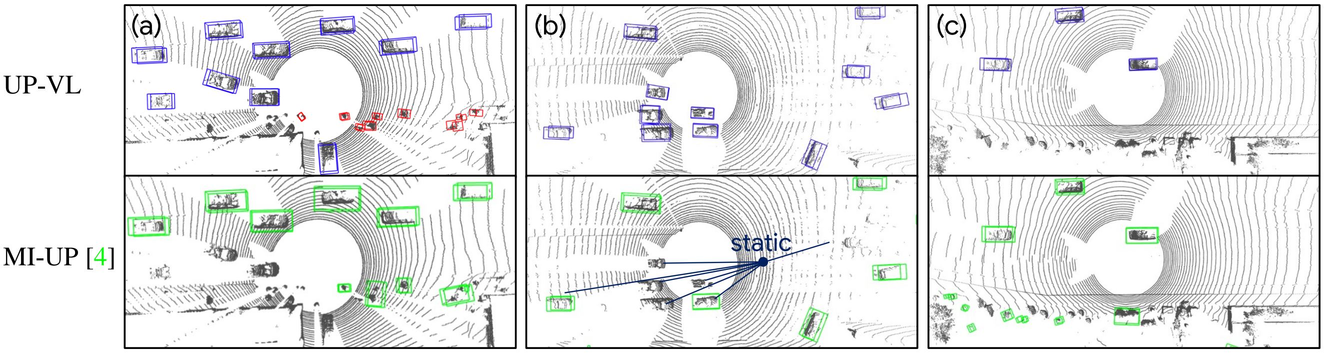

Figure S1. Comparison of our UP-VL with prior work MI-UP [4]. Comparatively, our UP-VL (a) localizes objects and classifies them, (b) detects both moving and static objects, (c) produces fewer false positives. Best viewed in color. Box colors: blue for vehicle, red for VRU, green for class-agnostic.

Table S1. Effect of hyperparameters of ϵ^{bg} and r^{bg}.

ϵ^{bg}	3D AP		mAP	r^{bg}	3D AP		mAP
	Veh	VRU			Veh	VRU	
0.10	28.7	12.3	20.5	50%	20.7	7.5	14.1
0.05	29.7	14.1	21.9	90%	27.3	11.1	19.2
0.02	**30.2**	**14.7**	**22.4**	**99%**	**30.2**	**14.7**	**22.4**
0.00	29.9	14.3	22.1	100%	30.1	14.6	22.3

ing moving-only objects by design, UP-VL is able to detect static objects as well. Lastly, by comparing column (c), one can see that our UP-VL approach can significantly reduce the false positives on cluttered parts of the scene, showing yet another advantage of our approach compared to the prior work on unsupervised 3D object detection in autonomous driving.

3. Effect of Hyperparameters

In this subsection, we perform an ablation study on the effect of the hyper-parameters introduced in Algorithm 1 of the main paper. More specifically, ϵ^{bg} which is used as a threshold on the computed cosine similarities to define the background points, and r^{bg} which represents a threshold on the required ratio of background points within a box proposal to mark it as background and consequently filtering the proposal. The ablation analysis is presented in Table S1. First thing to notice is that our approach is fairly robust to these hyper parameters when they are set in a reasonable range. Moreover, comparing the middle rows with the first and last rows demonstrates the effectiveness of introducing these thresholding schemes in improving the mAP of the model. Given these results, in all experiments in the paper we set $\epsilon^{bg} = 0.02$ and $r^{bg} = 0.99$.

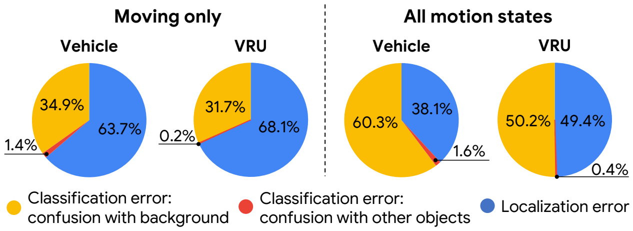

Figure S2. Error analysis of false positives. Fractions of false-positives that are caused by classification or localization errors. Our analysis covers two scenarios: detecting moving objects only and detecting objects in all motion states. And we examine both vehicle and VRU categories.

4. Error Analysis

4.1. Quantitative Analysis

Section 4 in the main paper discusses the overall accuracy of our open-vocabulary 3D object detectors. In this subsection, we will delve deeper into the analysis by breaking down the errors. One significant type of errors is false positive detections, which occurs when the detected object does not correspond to any ground truth object, given evaluation thresholds. Following Hoiem *et al.* [2], we categorize false positives into three types. **Localization error** arises when a detected object belongs to the intended category but has a misaligned bounding box ($0.1 <$ 3D IoU < 0.4). The remaining false positives, which have an IoU of at least 0.1 with an ground-truth object from a different category, are classified as **confusion with other objects**. All other false positives fall under the category of **confusion with background**. For each category, we count the "top-ranked" false positives among the most confident N detections, where N is selected to be half the quantity of ground truth objects in that category. Results are presented in Figure S2. It should be noted that given the decoupled design of our detector, the localization error can be linked to our class-agnostic bounding box prediction branch, and the classification error can be linked to our distillation branch. As can be seen, for moving objects (the left side of the figure), the localization error is the bottleneck in performance. This is while, when we also consider the static objects (the right side of the figure), the share of the classification error noticeably increases. Moreover, as expected, we can see that confusion between the categories (vehicles *vs*. VRUs) accounts for a very small portion of the false positives. We believe this analysis sheds light on the bottlenecks for further improvements of the proposed approach.

4.2. Qualitative Analysis

In the previous subsection, we performed quantitative error analysis on the available human annotations in the dataset. Here, we qualitatively present some error patterns of our method in the open-vocabulary setting where human annotations are unavailable. Figure S3 illustrates some real-world challenges in unsupervised open-vocabulary 3D detection. One type of failure case is the detector failing to generate a bounding box even though the point-wise cosine similarity has captured the correct semantics from the user's query (*e.g*. "tram" in Figure S3). We believe this is because such kind of large objects are rarely seen in the training data and our detector requires more unsupervised training data to confidently capture those objects. Another type of failure case is the mismatch between text queries and visual features for semantically similar concepts. Like the second example in Figure S3, where a text query of "truck" has matched with a crane. We hypothesize that this might be due to the similar appearance between cranes and construction trucks and the high co-occurrence of these two object types in the real-world.

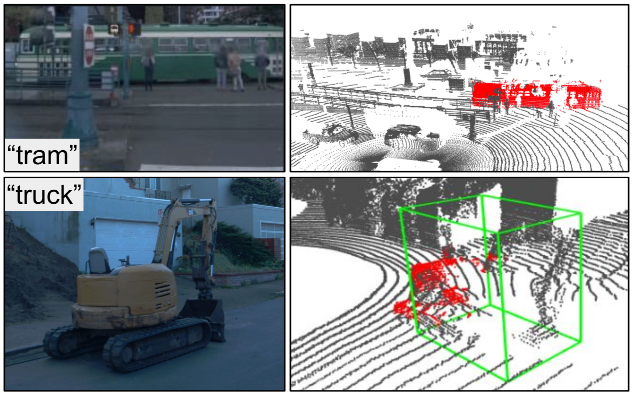

Figure S3. Failure cases. (a) Detector fails to generate very large boxes for rare categories like ”tram” although the point-wise semantic assignment is correct. (b) Text query of ”truck” wrongly matches with an object of crane.

References

[1] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 1

[2] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV (3)*, pages 340–353, 2012. 2

[3] 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*, 2021. 1

[4] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *ECCV*, 2022. 1, 2