Open Vocabulary Monocular 3D Object Detection

Supplementary Material

Sec. A discusses limitations of this paper. Sec. B presents per-category performance on novel classes for OVMono3D-GEO and OVMono3D-LIFT. Sec. C provides additional qualitative visualizations on various datasets and compares predictions between OVMono3D-GEO and OVMono3D-LIFT. Sec. D discusses failure cases, highlighting challenges such as occlusions, out-of-distribution objects, and small or distant instances. Sec. E provides more analysis on synthetic data and the naming ambiguity issue in current benchmarks.

A. Limitations

Due to the lack of 3D detection ground truth labels in COCO [35], the qualitative zero-shot evaluation is not feasible to perform. Additionally, our method requires accurate camera intrinsics as input; however, for in-the-wild images, the estimated intrinsics can be inaccurate, leading to errors in prediction. Furthermore, the use of computationally heavy components, such as Grounding DINO [36] and DINOv2 [43], results in slower inference speed compared to Cube R-CNN [3], which should be addressed in future work. See Appendix D for visualizations of failure cases.

B. Per-category Performance on novel classes

We show per-category performance on 3D Average Precision (AP_{3D}) for OVMono3D-GEO and OVMono3D-LIFT in Tab. 6.

C. More Qualitative Results

Additional qualitative visualizations of OVMono3D-LIFT are provided for Omni3D [3] outdoor, indoor subsets, and COCO [35] in-the-wild images in Figs. 8 to 10, respectively. For COCO images, we visualize with intrinsics of $f = 2 \cdot H$, $p_x = \frac{1}{2}W$, $p_y = \frac{1}{2}H$, where $H \times W$ is the input image resolution.

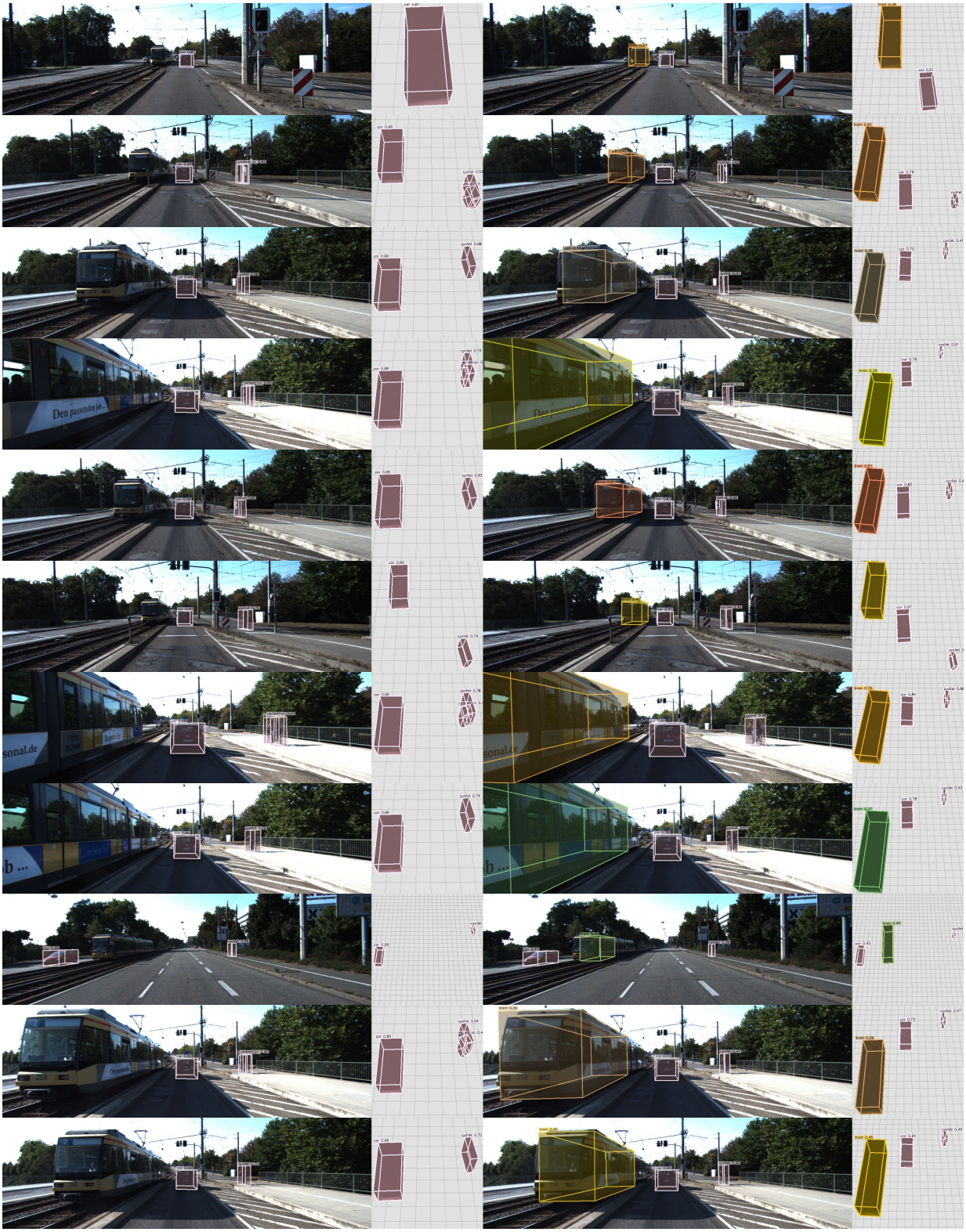
Fig. 11 illustrates a comparison between the predictions of OVMono3D-GEO and OVMono3D-LIFT. OVMono3D-GEO derives object depth from an estimated metric depth map, yielding better relative depth consistency with scene layout (*e.g.*, Fig. 11c). However, it estimates dimensions and poses based on visible object parts, leading to biases. For instance, it struggles with occlusions (*e.g.*, the door in Fig. 11d), limited surface visibility (*e.g.*, ovens in Fig. 11e), and noisy depth maps (*e.g.*, the farthest chair in Fig. 11f). In contrast, OVMono3D-LIFT, leveraging learned priors, is more robust in such scenarios. Future work could integrate these methods to mitigate their respective limitations.

D. Failure Cases

Fig. 12 shows failure cases of OVMono3D-LIFT on COCO [35] images. In Fig. 12a, the relative position from a top-down view is incorrect, indicating that our model sometimes predicts the wrong object depth. In Fig. 12b, our model fails to predict the correct size and pose for the bear, suggesting that it may struggle with totally out-of-distribution objects. In Fig. 12c, our model fails to detect the person and bus in the distance, indicating that it may not perform well on small and distant objects. In Fig. 12d, our method fails to identify the mirror and incorrectly detects the object in the mirror. These failure cases suggest that

Category	GEO	LIFT
Board	12.42	9.92
Printer	34.03	35.90
Painting	5.16	6.31
Microwave	33.47	48.54
Tray	10.35	15.56
Podium	41.74	62.18
Cart	32.22	54.00
Tram	0.25	8.65
Easy Categories	21.20	30.13
Monitor	18.25	13.92
Bag	25.48	23.96
Dresser	29.78	36.71
Keyboard	15.44	12.58
Drawers	30.01	57.21
Computer	13.14	13.77
Kitchen Pan	15.44	19.90
Potted Plant	20.13	6.07
Tissues	13.49	18.28
Rack	14.60	15.74
Toys	24.07	21.70
Phone	22.21	11.37
Soundsystem	17.73	17.69
Fireplace	24.44	19.76
Hard Categories	20.30	20.62
All Categories	20.63	24.08

Table 6. Per-category Performance of OVMono3D-GEO and OVMono3D-LIFT. The reported metric is AP_{3D} in target-aware evaluation.



Cube R-CNN OVMono3D-LIFT

Figure 8. **Qualitative Visualizations on the KITTI [15] Test Set.** For each example, we present the predictions of Cube R-CNN [3] and OVMono3D-LIFT, displaying both the 3D predictions overlaid on the image and 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.

Figure 9. **Qualitative Visualizations on the SUN RGB-D [59] Test Set.** For each example, we present the predictions of Cube R-CNN [3] and OVMono3D-LIFT, displaying both the 3D predictions overlaid on the image and a top-down view with a base grid of 1 m × 1 m tiles. Base categories are depicted with brown cubes, while novel categories are represented in other colors.

our model still has room for improvement. Future research could explore better model architectures and weakly supervised learning techniques to address these shortcomings.

E. More Analysis

Do synthetic data help? We conducted an ablation study using synthetic data for OVMono3D-LIFT under resource-constrained conditions, with a frozen image encoder and excluded depth estimator. The synthetic data comes from Hypersim [53], which provides indoor images rendered from artist-created meshes and serves as the sole synthetic data source in Omni3D [3].

Tab. 7 presents the effect of synthetic data on the performance of OVMono3D-LIFT. When synthetic data is incorporated alongside real data, a modest yet meaningful increase of 1 AP_{3D} point is observed in detecting objects from seen categories, while performance on novel categories remains largely unchanged. These findings suggest that, while synthetic data can enhance model performance in closed-vocabulary 3D object detection tasks, its benefits are minimal for detecting unseen objects, thereby limiting its usefulness in open-vocabulary 3D object detection sce-

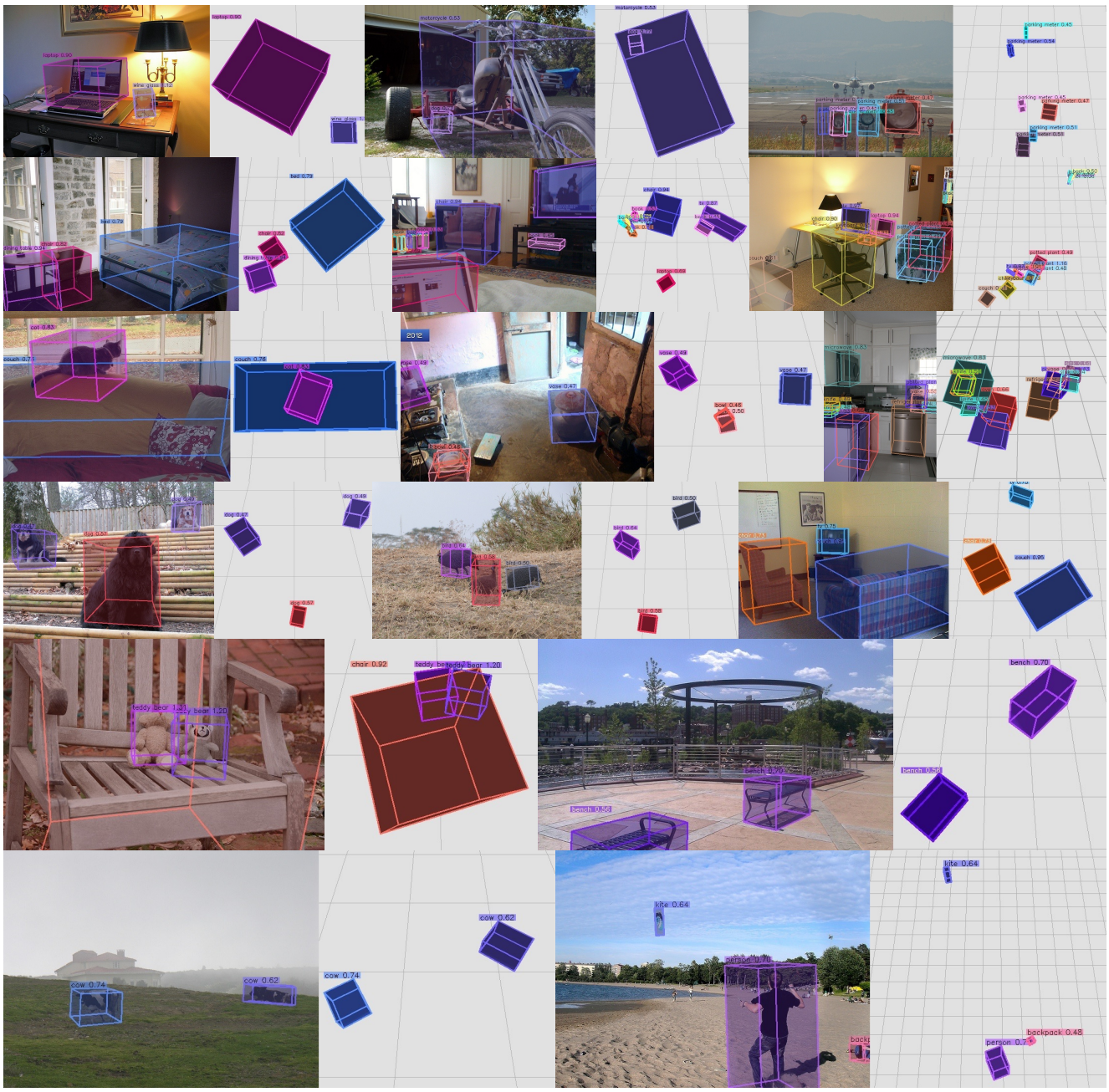


Figure 10. **OVMono3D-LIFT on In-the-Wild COCO [35] Images**. We display 3D predictions overlaid on the images and the top-down views with a base grid of 1 m × 1 m tiles.

narios.

Naming Ambiguity issue. We quantitatively evaluate naming ambiguity in current 3D benchmarks using SUN RGB-D [59] as an example. For each object instance, we cropped its 2D bounding box and computed CLIP similarity scores between the visual features and all category text embeddings. We then aggregated these similarity vectors by ground-truth category and computed average similarities to form a confusion matrix, applying softmax normalization.

As shown in Fig. 13, SUN RGB-D annotations exhibit weaker self-correlation than COCO [35], indicating less distinct category boundaries. This reflects SUN RGB-D's highly similar category names (e.g., "table" vs. "desk"). In open-vocabulary settings, such similarity creates false negatives when models correctly identify a table as a

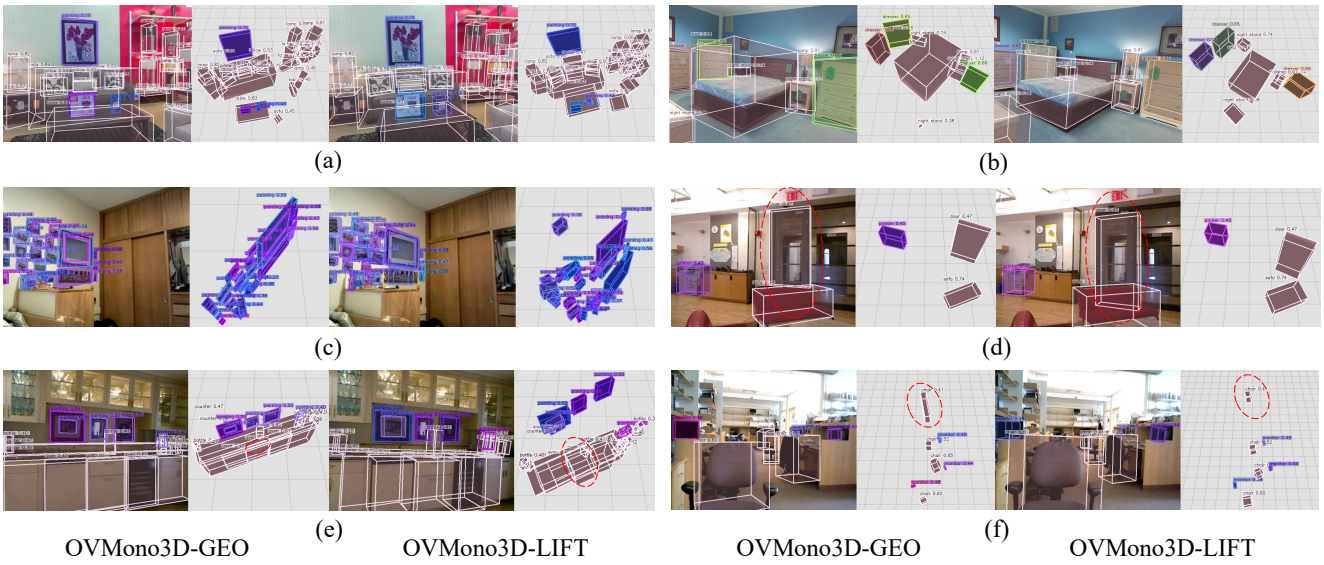


Figure 11. **OVMono3D-GEO *vs.* OVMono3D-LIFT on SUN RGB-D [59] Images**. For each example, we display the predictions of OVMono3D-GEO and OVMono3D-LIFT. We display 3D predictions overlaid on the images and the top-down views with a base grid of 1 m $\times$ 1 m tiles. Base categories are depicted with brown cubes, while novel categories are represented in other colors.

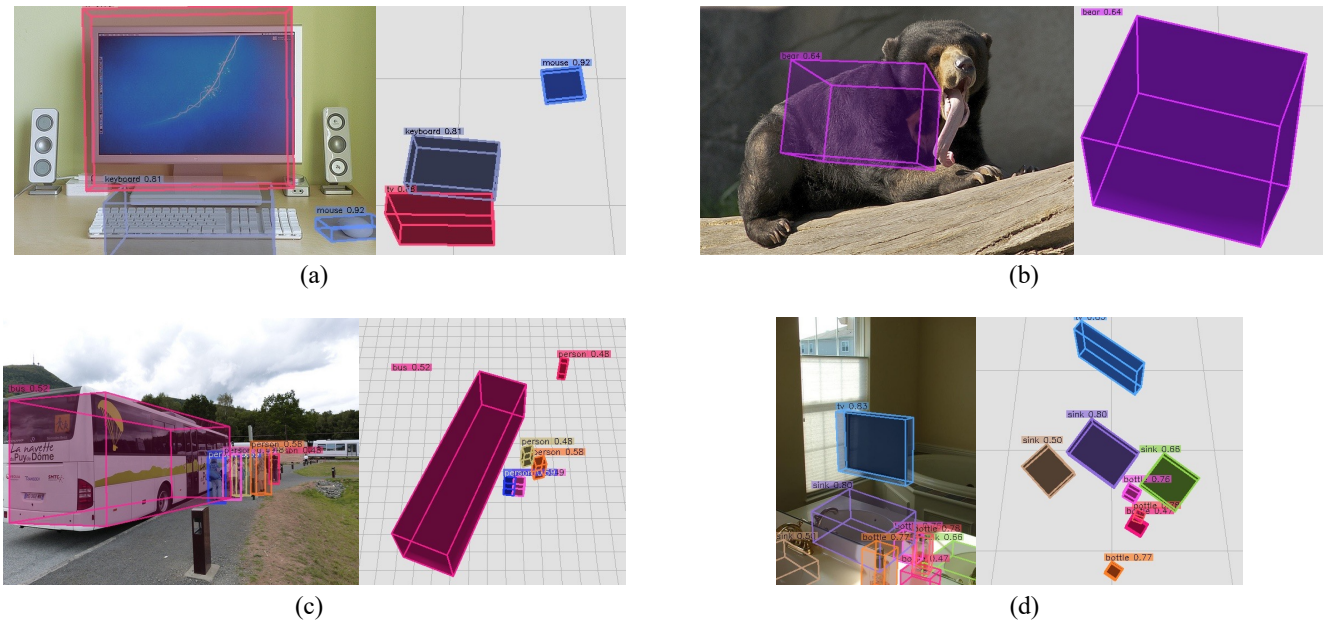


Figure 12. **Failure Cases of OVMono3D-LIFT on COCO [35] Images**. We display 3D predictions overlaid on the images and the top-down views with a base grid of 1 m $\times$ 1 m tiles.

desk—a distinction often acceptable in real-world applications. Therefore, our proposed target-aware evaluation is essential for datasets with ambiguous category definitions, unlike the well-differentiated categories in COCO.

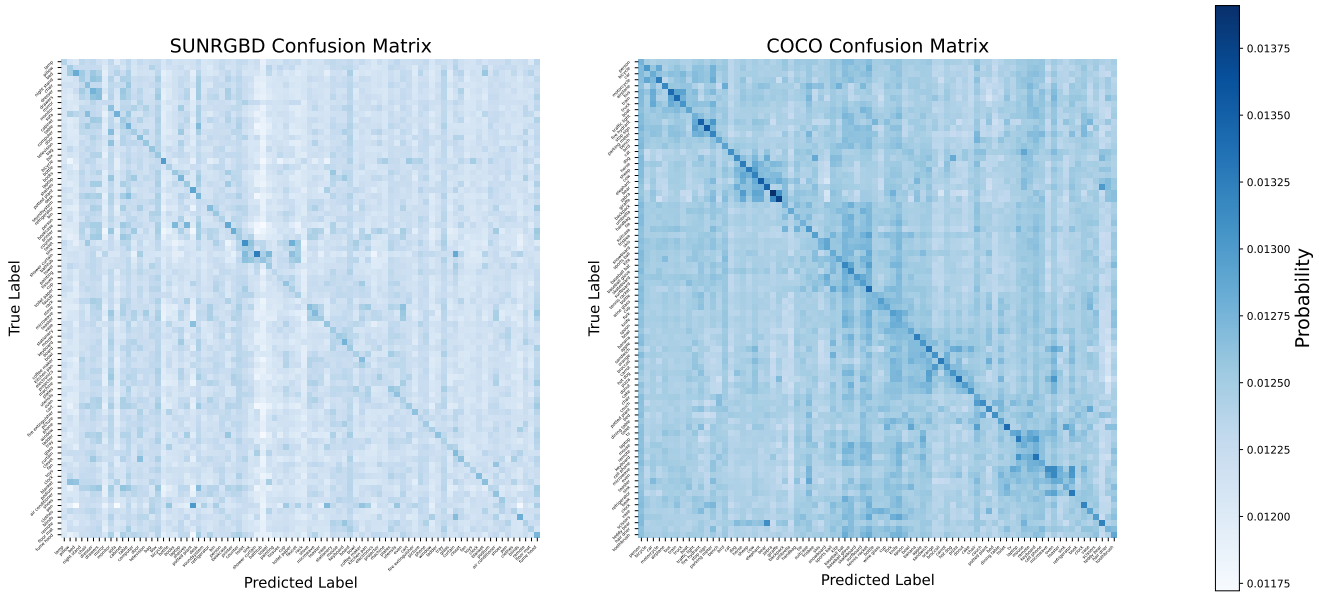


Figure 13. Normalized confusion matrices displaying CLIP's prediction performance on the SUNRGBD and COCO datasets.

Data	#Images	AP_{3D}^{Base}	AP_{3D}^{Novel}
Synthetic	55k	7.14	7.33
Real	120k	23.78	**16.20**
Synthetic+Real	175k	**24.77**	16.04

Table 7. **Ablation on Synthetic Data.** Synthetic data refers to the Hypersim subset of the Omni3D dataset, while real data comprises the other Omni3D subsets. Synthetic data boosts the performance of OVMono3D-LIFT on base categories but offers little benefit for novel objects.