3D-MOOD: Lifting 2D to 3D for Monocular Open-Set Object Detection

Supplementary Material

This supplementary material elaborates more details of our main paper. In Sec. A, we illustrate how the proposed canonical image space can reduce the required GPU resource for training and how it helps the model learn better prior geometry. In Sec. B, we compare our proposed geometry-aware query generation to the virtual depth proposed by Cube R-CNN [1]. In Sec. C, we provide further details of our proposed open-set benchmark. We analyze the depth estimation performance in Sec. D, backbone comparison in Sec. E and FPS in Sec. F, respectively. In Sec. G, we discuss the importance of our proposed open detection score (ODS) and compare the evaluation results in detail to the IoU-based AP. Finally, we provide more qualitative results in Sec. H for both closed-set and open-set settings.

A. Canonical Image Space

As shown in the main paper, we compare different resizing and padding strategies for training the model. Given the training batch size as 2, we have two training samples having very different image ratios, *e.g.* $[376 \times 1241]$ and $[1920 \times 1080]$. The first strategy as [1] will find the shortest edge and resize it to the desired value, *e.g.* 512, and conduct the right-bottom padding to align the two samples' resolutions. This will lead to considerable padding for the portrait image, while not change the camera intrinsic $\mathbf{K}$.

The second strategy is like Grounding DINO (G-DINO) [4], which will find the longest edge and resize it to the desired value, *e.g.* 1333, and use the same right-bottom padding to align the resolutions. This leads to considerable padding for the landscape image, while the padding will also not change the camera intrinsic $\mathbf{K}$. Both strategies will increase the GPU usage for unnecessary padding and lead to different image resolutions for the same sampled image according to the paired images.

On the other hand, our proposed canonical image space fixes the image resolutions, *e.g.* $[800 \times 1333]$, and will resize the longest or the shortest edge considering the image ratios. As shown in Tab. 5, our methods successfully reduce the needed GPU resources compared to the previous methods. Furthermore, we use the center padding to ensure our image space will affect the camera intrinsic $\mathbf{K}$ accordingly to unify it across not only the training and testing time but also across datasets.

During inference time, the same camera will capture the same image shape with the same camera intrinsics. The previous methods will fail to align the same observation between training and inference time. We speculate that this will hinder the model's understanding of the relation between intrinsics, image shape, and metric depth. On the contrary, our canonical image space will keep the image and intrinsic consistent. As shown in the ablation studies of the main paper, the model benefits from the proposed canonical image space for both closed-set and open-set settings with even fewer GPU resource requirements for training.

Table 5. **GPU RAM Consumption**. We compare the GPU resource usage of different training padding and resizing methods. We show the results of training our full model using Swin-T [5] and batch size of 2 using gradient checkpointing on a RTX 4090.

Resize	Padding	Image Resolutions	GPU RAM (G)
Short Edge	Right-Bottom	Short Edge to 512	21
Long Edge	Right-Bottom	Long Edge to 1333	23
Ours	Center	800×1333	17

Table 6. **Comparison with Virtual Depth**. We compare our geometry-aware 3D query generation (GA) with the virtual depth proposed by Cube R-CNN [1]. The results shows that GA converge better than the virtual depth mechanism.

Method	Virtual Depth	GA	AP_{3D}^{omni} ↑
Cube R-CNN [1]	✓	-	23.3
Ours (Swin-T)	✓	-	21.6
Ours (Swin-T)	-	✓	**26.8**

B. Comparison with Virtual Depth

We compare our proposed geometry-aware 3D query generation with the virtual depth proposed in [1]. As shown in Tab. 6, the virtual depth leads our model to converge much slower than our proposed geometry-aware 3D query generation. We speculate that virtual depth requires much more training time to learn universal geometry, which also leads to underperformance.

C. Open-set Benchmark

We show more details about our proposed open-set benchmark and list the full classes for each datasets in Tab. 7.

C.1. Argoverse 2

Compared to other autonomous driving datasets, Argoverse 2 (AV2) [10] possesses more diverse classes. Moreover, the resolutions of the front camera are portrait images, providing unseen cameras and domains. Those factors make AV2 a challenging dataset for benchmarking open-set monocular 3D object detection. We sample every 5 frame from the official validation split and obtain 4806 images as the open-set

Table 7. **Classes for Argoverse 2 and ScanNet.** We list the base and novel categories for our proposed open-set benchmark. For ScanNet, the bold categories are the supercategories. We further list all 168 categories of ScanNet200 settings.

Dataset	Base	Novel
Argoverse 2	regular vehicle, pedestrian, bicyclist, construction cone, construction barrel, large vehicle, bus, truck, vehicular trailer, bicycle, motorcycle	motorcyclist, wheeled rider, bollard, sign, stop sign, box truck, articulated bus, mobile pedestrian crossing sign, truck cab, school bus, wheeled device, stroller
ScanNet	**cabinet** (cabinet, kitchen cabinet, file cabinet, bathroom vanity, bathroom cabinet, cabinet door, trash cabinet, media center), **bed** (bed, mattress, loft bed, sofa bed, air mattress), **chair** (chair, office chair, armchair, sofa chair, folded chair, massage chair, recliner chair, rocking chair), **sofa** (couch, sofa), **table** (table, coffee table, end table, dining table, folded table, round table, side table, air hockey table), **door** (door, doorframe, bathroom stall door, closet door, mirror door, glass door, sliding door, closet doorframe), **window**, **picture** (picture, poster, painting), **counter** (kitchen counter, counter, bathroom counter), **desk**, **curtain**, **refrigerator** (refrigerator, mini fridge, cooler), **toilet** (toilet, urinal), **sink**, **bathtub**	**bookshelf**, **shower curtain**, **other furniture** (trash can, radiator, recycling bin, ottoman, bench, tv stand, wardrobe, trash bin, seat, closet, ladder, piano, water cooler, stand, washing machine, rack, wardrobe , clothes dryer, ironing board, keyboard piano, music stand, furniture, crate, drawer, footrest, piano bench, foosball table, footstool, compost bin, tripod, treadmill, chest, folded ladder, drying rack, pool table, heater, toolbox, beanbag chair, dollhouse, ping pong table, clothing rack, podium, luggage stand, rack stand, futon, book rack, workbench, easel, headboard, display rack, crib, bedframe, bunk bed, magazine rack, furnace, stepladder, baby changing station, flower stand, display)
ScanNet200	chair, table, door, couch, cabinet, shelf, desk, office chair, bed, pillow, sink, picture, window, toilet, bookshelf, monitor, curtain, book, armchair, coffee table, box, refrigerator, lamp, kitchen cabinet, towel, clothes, tv, nightstand, counter, dresser, stool, plant, bathtub, end table, dining table, keyboard, bag, backpack, toilet paper, printer, tv stand, whiteboard, blanket, shower curtain, trash can, closet, stairs, microwave, stove, shoe, computer tower, bottle, bin, ottoman, bench, board, washing machine, mirror, copier, basket, sofa chair, file cabinet, fan, laptop, shower, paper, person, paper towel dispenser, oven, blinds, rack, plate, blackboard, piano, suitcase, rail, radiator, recycling bin, container, wardrobe, soap dispenser, telephone, bucket, clock, stand, light, laundry basket, pipe, clothes dryer, guitar, toilet paper holder, seat, speaker, column, ladder, cup, jacket, storage bin, coffee maker, dishwasher, paper towel roll, machine, mat, windowsill, bar, bulletin board, ironing board, fireplace, soap dish, kitchen counter, doorframe, toilet paper dispenser, mini fridge, fire extinguisher, ball, hat, shower curtain rod, water cooler, paper cutter, tray, pillar, ledge, toaster oven, mouse, toilet seat cover dispenser, cart, scale, tissue box, light switch, crate, power outlet, decoration, sign, projector, closet door, vacuum cleaner, headphones, dish rack, broom, range hood, hair dryer, water bottle, vent, mailbox, bowl, paper bag, projector screen, divider, laundry detergent, bathroom counter, stick, bathroom vanity, closet wall, laundry hamper, bathroom stall door, ceiling light, trash bin, dumbbell, stair rail, tube, bathroom cabinet, coffee kettle, shower head, case of water bottles, power strip, calendar, poster, mattress	

testing set. Among the official 26 classes, 23 appeared in the testing set, which contains 11 *base* and 12 *novel* classes.

C.2. ScanNet

ScanNet [2] provides diverse indoor scenes with 18 supercategories as shown in Tab. 7. We uniformly sample maximum 20 frames from each scan in the official ScanNet validation splits and obtain total 6240 images as open-set testing set. Given that 15 supercategories are seen in the Omni3D training set, this benchmark still allows us to evaluate domain generalization, where Tab. 4 of the main paper indicates issues of previous works. Furthermore, in the rest 3 novel classes, the supercategory *other furniture* requires models to detect various types of furniture.

To further test 3D-MOOD, we extend ScanNet using the ScanNet200 setting, which has **168** thing classes appeared in the testing set. As shown in Tab. 8, 3D-MOOD can still achieve best performance given more diverse classes.

Table 8. **ScanNet200 Results.** 3D-MOOD achieves SOTA results given diverse novel categories in unseen scenes.

Method	AP_{3D}^{dist} ↑	mATE ↓	mASE ↓	mAOE ↓	ODS ↑
Cube R-CNN [1]	2.1	0.962	0.970	0.985	2.5
OVM3D-Det [3]	3.1	0.957	0.973	0.946	3.6
Ours (Swin-B)	**6.2**	**0.811**	**0.835**	**0.799**	**12.4**

D. Metric Monocular Depth Estimation

Because auxiliary depth estimation (ADE) is only used to help with 3D object detection, we evaluated its effectiveness in this regard. As shown in Tab. 4 of our paper, ADE improves the closed set AP by 0.7, but reduces the performance for unseen scenes, indicating that ADE can only help for known scenes. We further evaluate our depth quality on the KITTI Eigen-split test set, where UniDepth [8] achieves 4.21% absolute relative error, Metric3Dv2 [11] has

Table 9. **Backbone comparison.** We ablate the choice of different model backbones. All experiments are trained with 12 epochs.

Backbone	Parameters	AP_{3D}^{omni} ↑
DLA-34 [12]	15M	24.9
Swin-Transformer (Tiny) [5]	29M	26.8
Swin-Transformer (Base) [5]	88M	28.2
ConvNeXt-B [6]	89M	28.4

Table 10. **Comparison between different matching criteria for evaluation.** The same detection results will have a huge AP difference when using different matching settings.

Matching	Pedestrian	Construction cone	Monitor	Door
IoU	7.4	0.5	0.5	2.0
Distance	26.2	6.5	9.4	24.2

4.4%, and 3D-MOOD obtains 9.1%. We believe it is due to limited training data, different training objectives, and the model backbones [5, 7].

E. Backbone Comparison

As shown in Tab. 9, Swin-B works equally well with ConvNeXt-B, and 3D-MOOD is comparable to Cube R-CNN and Uni-MODE using the same backbone, but with much shorter training. This shows effectiveness of our proposed designs rather than the backbone [7].

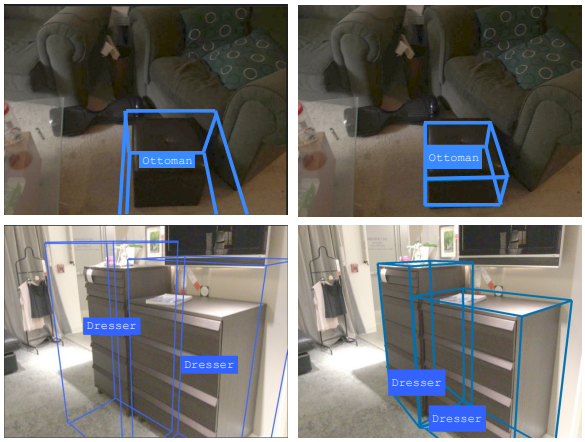
F. Inference Time

We compare the FPS on KITTI using an RTX 4090, and Cube R-CNN (DLA-34) can have 68 FPS while 3D-MOOD (Swin-T) can achieve 17 FPS. As a reference from the paper, Uni-MODE can obtain 21 FPS on a single A100.

G. Open Detection Score (ODS)

Compared to point-cloud-based 3D object detectors, using a single image to estimate 3D objects requires the networks to predict metric depth, while the scales in depth are known in the point cloud. This extra challenge leads the monocular methods to fail to match the ground truth using IoU-based matching because of several centimeter error in depth, especially for the small or thin objects in the open-set settings.

To have a more suitable evaluation metric for monocular 3D object detection, we use the 3D Euclidean distance between prediction and GT as the matching criterion. With the dynamic matching threshold, *e.g. radius* of the GT 3D boxes, AP_{3D}^{dist} can be used for both indoor and outdoor scenes. As shown in Tab. 10, the same detection results on Argoverse 2 [10] and ScanNet [2] will have large AP differences depending on the matching criterion.



Gemini 2 **3D-MOOD**

Figure 6. **Comparison with Gemini 2.** We qualitatively compare with Gemini 2 given the novel classes.

We also propose the normalized true positive errors (TPE) to further analyze the matched prediction. First, we compute the 3D Euclidean distance between prediction and GT, and we normalize the distance by the matching criterion as the translation error (TE). Second, we compute the IoU, *i.e.* IoU_{3D}, between prediction and GT after aligning the 3D centers and orientation and use $1-IoU_{3D}$ to measure the scale error (SE). Finally, we compute the SO3 relative angle between the prediction and GT normalized by π as the orientation error (OE). We average the TP errors across classes over different recall thresholds to get mATE, mASE, and mAOE. Using AP_{3D}^{dist} with the proposed normalized true positive errors to get ODS can provide a better matching criterion for 3D monocular object detection and still evaluate the localization, orientation, and dimension estimation at the same time.

H. Qualitative Results

As shown in Fig. 6, we qualitatively compared our method with the closed-source Gemini 2 [9] beta functionality in 3D object detection, where 3D-MOOD provides more accurate localization. We provide more qualitative results in Fig. 7 for the open-set settings and Fig. 8 for the closed-set settings. We use the score threshold as 0.1 with class-agnostic nonmaximum suppression for better visualization.

References

[1] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 1, 2, 5

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Hal-

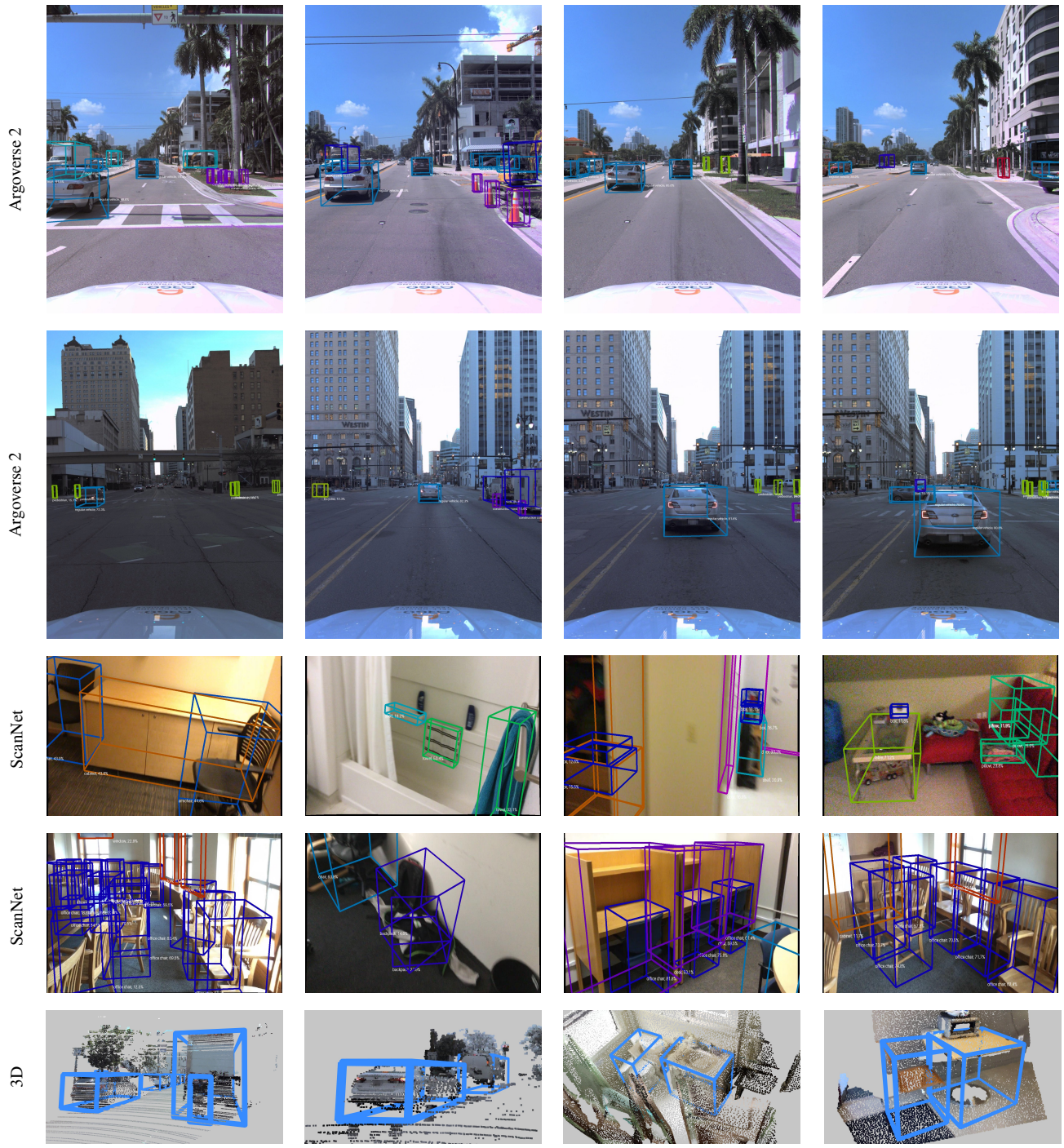


Figure 7. **Open-set Qualitative Results.** We show more visualization on Argoverse 2 [10] and ScanNet [2].

ber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 3, 4

[3] 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. 2

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao

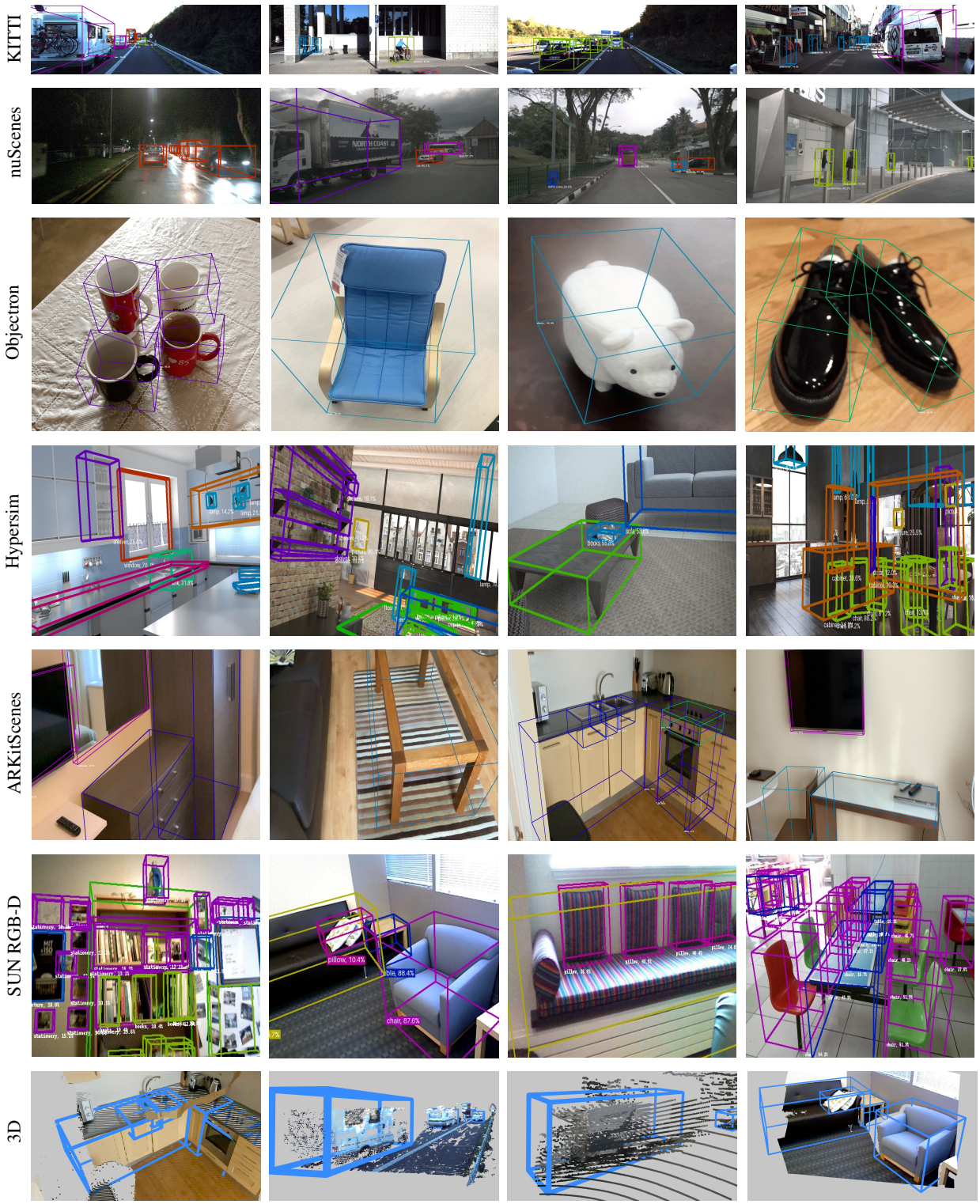


Figure 8. **Closed-set Qualitative Results.** We show the qualitative results for 3D-MOOD on Omni3D [1] test set.

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. 1

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 3

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[7] 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. 3

[8] 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 (CVPR)*, 2024. 2

[9] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[10] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 1, 3, 4

[11] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 2

[12] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 3