Geometry-based Distance Decomposition for Monocular 3D Object Detection
Supplementary File

	AP_{3D} / AP_{BEV} (0.5) ↑	AP_{3D} / AP_{BEV} (0.7) ↑
M3D-RPN [1]	17.50 / 20.40	5.12 / 9.51
Ours	23.66 / 26.83	8.15 / 12.64

Table 1. AP_{3D} / AP_{BEV} (IoU >0.5 and 0.7) on the nuScenes val subset [2].

1. KITTI Val Examples

We show qualitative examples of our MonoRCNN and M3D-RPN [1] on the val subset of the KITTI val split [3] in Fig. 1. The results show our method is more accurate.

2. Cross-Dataset Test Examples

We show qualitative examples of our MonoRCNN and M3D-RPN [1] on the nuScenes [2] cross-test set in Fig. 2. We can see our method is more accurate.

3. nuScenes Results

To further show our generalizability, we train and evaluate our method on the nuScenes dataset [2]. Following [4], we use the front camera and consider objects in its FOV, and evaluate on the val subset. We extract the images and labels of the front camera with a nuScenes official KITTI converter [1]. There are $28\,130$ training images (about 4 times larger than KITTI). Following [4], we train a model of M3D-RPN [1] using its official code for a comparison. We report AP_{3D} and AP_{BEV} for cars under IoU criteria 0.5 and 0.7 using KITTI official evaluation tool. We observe : 1). The mean prediction error of the physical length, width, and height of cars on the nuScenes val subset are 0.283m, 0.128m, and 0.118m, respectively, supporting that the physical height is the easiest variable among physical size. 2) For our model trained on nuScenes, its AP_{3D} / AP_{BEV} on the nuScenes val subset decreases, if predicted H for recovering the distance is replaced with the groundtruth H. This decrease shows that the correlation between predicted H and h_{rec} also exists on nuScenes. These two observations on nuScenes are consistent with on KITTI and beneficial to distance estimation. As shown in Tab. 1, ours outperforms M3D-RPN [1] by a large margin.

[1] https://github.com/nutonomy/nuscenes-devkit/blob/master/python-sdk/nuscenes/scripts/export_kitti.py

References

[1] Garrick Brazil and Xiaoming Liu. M3D-RPN: monocular 3d region proposal network for object detection. In *ICCV*, 2019.

[2] 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] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G. Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *NeurIPS*, 2015.

[4] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021.

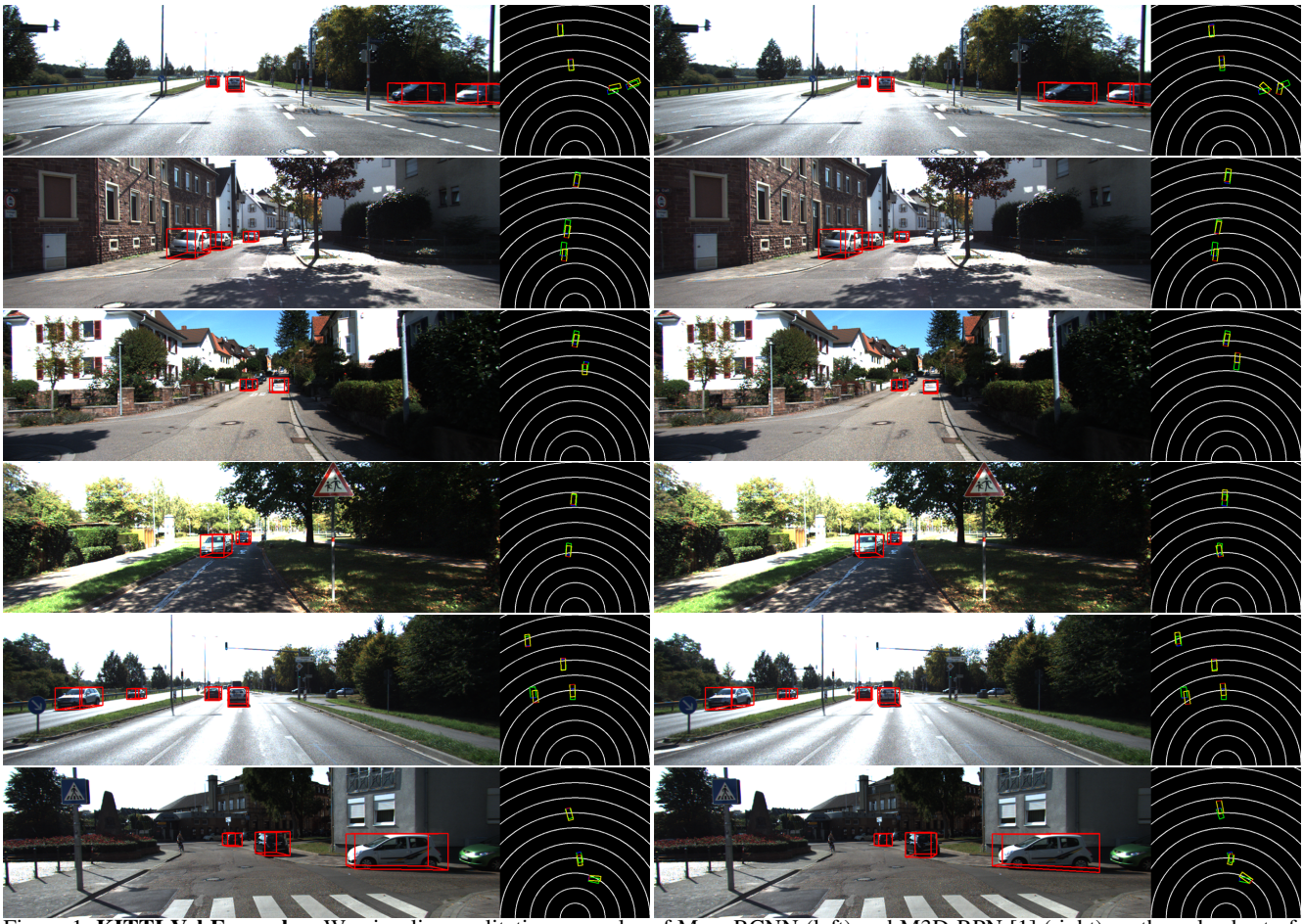

Figure 1. **KITTI Val Examples**. We visualize qualitative examples of MonoRCNN (left) and M3D-RPN [1] (right) on the val subset of the KITTI val split [3]. We can see our method is more accurate than M3D-RPN [1]. The red boxes in the image planes represent the 2D projections of the predicted 3D bounding boxes. The yellow / green boxes in the bird's eye view results represent the predictions and groundtruths of the 3D bounding boxes, respectively, and the red / blue lines indicate the yaw angle of cars. The radius difference between two adjacent white circles is 5 meters. All illustrated images are not used for training.

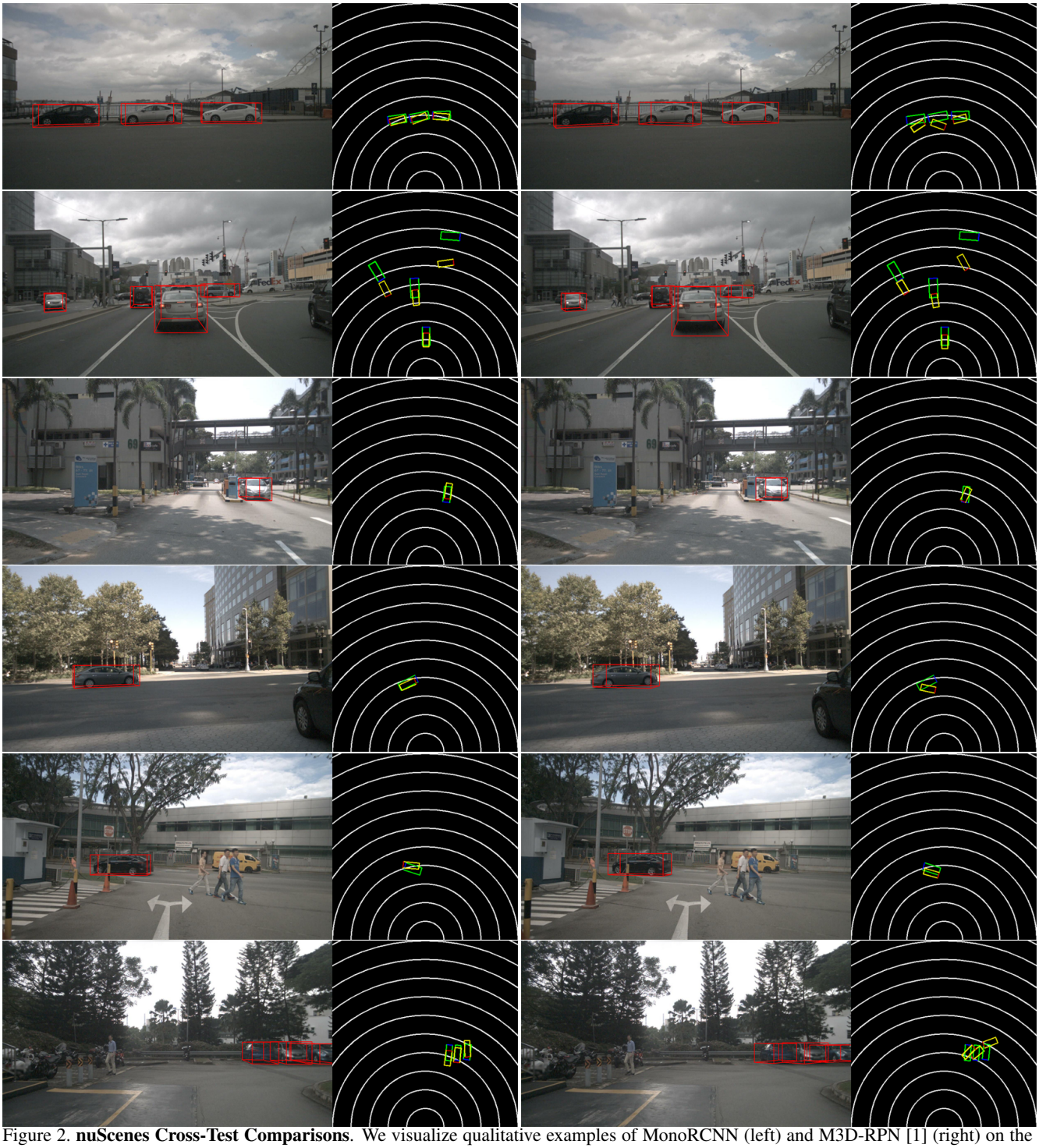

Figure 2. **nuScenes Cross-Test Comparisons**. We visualize qualitative examples of MonoRCNN (left) and M3D-RPN [1] (right) on the nuScenes [2] cross-test set. We can see our method achieves more accurate distance prediction. The 2D projections and bird's eye view results are shown as in Fig. 1. All models are only trained with the training subset of the KITTI val split [3].