Supplementary Material: Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger
Cornell University, Ithaca, NY
{yw763, wc635, dg595, bh497, mc288, kqw4}@cornell.edu

In this Supplementary Material, we provide details omitted in the main text.

- Section A: additional details on our approach (Section 4.2 of the main paper).
- Section B: results using SPS-STEREO [11] (Section 4.3 of the main paper).
- Section C: further analysis on depth estimation (Section 4.3 of the main paper).
- Section D: additional results on the test set (Section 4.4 of the main paper).
- Section E: additional qualitative results (Section 4.5 of the main paper).

A. Additional Details of Our Approach

A.1. Ground plane estimation

As mentioned in the main paper, AVOD [6] takes image-specific ground planes as inputs. A ground plane is parameterized by a normal vector $\boldsymbol{w} = [w_x, w_y, w_z]^\top \in \mathbb{R}^3$ and a ground height $h \in \mathbb{R}$. We estimate the parameters according to the pseudo-LiDAR points $\{\boldsymbol{p}^{(n)} = [x^{(n)}, y^{(n)}, z^{(n)}]^\top\}_{n=1}^N$ (see Section 3 of the main paper). Specifically, we consider points that are close to the camera and fall into a certain range of possible ground heights:

$$\text{(width)} \quad 15.0 \geq x \geq -15.0, \tag{1}$$
$$\text{(height)} \quad 1.86 \geq y \geq 1.5, \tag{2}$$
$$\text{(depth)} \quad 40.0 \geq z \geq 0.0. \tag{3}$$

Ideally, all these points will be on the plane: $\boldsymbol{w}^\top \boldsymbol{p} + h = 0$. We fit the parameters with a straight-forward application of RANSAC [2], in which we constraint $w_y = -1$. We then normalize the resulting $\boldsymbol{w}$ to have a unit ℓ_2 norm.

Table A: Comparison of different stereo disparity methods on pseudo-LiDAR-based detection accuracy with AVOD. We report AP_{BEV} / AP_{3D} (in %) of the **moderate car** category at IoU = 0.7.

Method	Disparity	AP_{BEV} / AP_{3D}
AVOD	SPS-STEREO	39.1 / 28.3
	DISPNET-S	36.3 / 27.0
	DISPNET-C	36.5 / 26.2
	PSMNET	39.2 / 27.4
	PSMNET⋆	56.8 / 45.3

A.2. Pseudo disparity ground truth

We train a version of PSMNET [1] (named PSMNET⋆) using the 3,712 training images of detection, instead of the 200 KITTI stereo images [4, 8]. We obtain pseudo disparity ground truth as follows: We project the corresponding LiDAR points into the 2D image space, followed by applying Eq. (1) of the main paper to derive disparity from pixel depth. If multiple LiDAR points are projected to a single pixel location, we randomly keep one of them. We ignore those pixels with no depth (disparity) in training PSMNET.

B. Results Using SPS-STEREO [11]

In Table A, we report the 3D object detection accuracy of pseudo-LiDAR with SPS-STEREO [11], a non-learning-based stereo disparity approach. On the leaderboard of KITTI stereo 2015, SPS-STEREO achieves 3.84% disparity error, which is worse than the error of 1.86% by PSMNET but better than 4.32% by DISPNET-C. The object detection results with SPS-STEREO are on par with those with PSMNET and DISPNET, even if it is not learning-based.

C. Further Analysis on Depth Estimation

We study how over-smoothing the depth estimates would impact the 3D object detection accuracy. We train AVOD [6] and F-POINTNET [9] using pseudo-LiDAR with PSMNET⋆. During evaluation, we obtain over-smoothed

Table B: The impact of over-smoothing the depth estimates on the 3D detection results. We evaluate pseudo-LiDAR with PSMNET⋆. We report AP_{BEV} / AP_{3D} (in %) of the **moderate car** category at IoU = 0.7 on the validation set.

	Detection algorithm	
Depth estimates	AVOD	F-POINTNET
Non-smoothed	56.8 / 45.3	51.8 / 39.8
Over-smoothed	53.7 / 37.8	48.3 / 31.6

depth estimates using an average kernel of size 11×11 on the depth map. Table B shows the results: over-smoothing leads to degraded performance, suggesting the importance of high quality depth estimation for accurate 3D object detection.

D. Additional Results on the Test Set

We report the results on the pedestrian and cyclist categories on the KITTI test set in Table C. For F-POINTNET which takes 2D bounding boxes as inputs, [9] does not provide the 2D object detector trained on KITTI or the detected 2D boxes on the test images. Therefore, for the car category we apply the released RRC detector [10] trained on KITTI (see Table 5 in the main paper). For the pedestrian and cyclist categories, we apply Mask R-CNN [5] trained on MS COCO [7]. The detected 2D boxes are then inputted into F-POINTNET [9]. We note that, MS COCO has no cyclist category. We thus use the detection results of bicycles as the substitute.

On the pedestrian category, we see a similar gap between pseudo-LiDAR and LiDAR as the validation set (cf. Table 4 in the main paper). However, on the pedestrian category we see a drastic performance drop by pseudo-LiDAR. This is likely due to the fact that cyclists are relatively uncommon in the KITTI dataset and the algorithms have over-fitted. For F-POINTNET, the detected bicycles may not provide accurate heights for cyclists, which essentially include riders and bicycles. Besides, the detected bicycles without riders are false positives to cyclists, hence leading to a much worse accuracy.

We note that, so far no image-based algorithms report 3D results on these two categories on the test set.

E. Additional Qualitative Results

E.1. LiDAR vs. pseudo-LiDAR

We include in Fig. A more qualitative results comparing the LiDAR and pseudo-LiDAR signals. The pseudo-LiDAR points are generated by PSMNET⋆. Similar to Fig. 1 in the main paper, the two modalities align very well.

Table C: 3D object detection results on the **pedestrian** and **cyclist** categories on the *test* set. We compare pseudo-LiDAR with PSMNET⋆ (in blue) and LiDAR (in gray). We report AP_{BEV} / AP_{3D} at IoU = 0.5 (the standard metric). †: Results on the KITTI leaderboard.

Method	Input signal	Easy	Moderate	Hard
Pedestrian				
AVOD	Stereo	27.5 / 25.2	20.6 / 19.0	19.4 / 15.3
F-POINTNET	Stereo	31.3 / 29.8	24.0 / 22.1	21.9 / 18.8
AVOD	†LiDAR + Mono	58.8 / 50.8	51.1 / 42.8	47.5 / 40.9
F-POINTNET	†LiDAR + Mono	58.1 / 51.2	50.2 / 44.9	47.2 / 40.2
Cyclist				
AVOD	Stereo	13.5 / 13.3	9.1 / 9.1	9.1 / 9.1
F-POINTNET	Stereo	4.1 / 3.7	3.1 / 2.8	2.8 / 2.1
AVOD	†LiDAR + Mono	68.1 / 64.0	57.5 / 52.2	50.8 / 46.6
F-POINTNET	†LiDAR + Mono	75.4 / 72.0	62.0 / 56.8	54.7 / 50.4

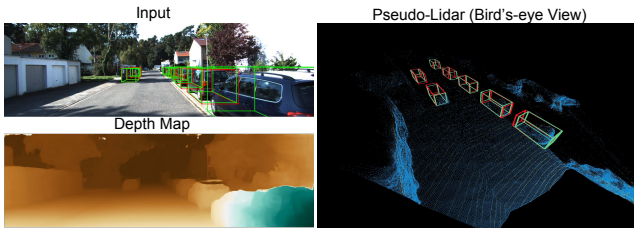


Figure A: **Pseudo-LiDAR signal from visual depth estimation.** Top-left: a KITTI street scene with super-imposed bounding boxes around cars obtained with LiDAR (red) and pseudo-LiDAR (green). Bottom-left: estimated disparity map. Right: pseudo-LiDAR (blue) vs. LiDAR (yellow) — the pseudo-LiDAR points align remarkably well with the LiDAR ones. Best viewed in color (zoom in for details).

E.2. PSMNET vs. PSMNET⋆

We further compare the pseudo-LiDAR points generated by PSMNET⋆ and PSMNET. The later is trained on the 200 KITTI stereo images with provided denser ground truths. As shown in Fig. B, the two models perform fairly similarly for nearby distances. For far-away distances, however, the pseudo-LiDAR points by PSMNET start to show notable deviation from LiDAR signal. This result suggest that significant further improvements could be possible through learning disparity on a large training set or even end-to-end training of the whole pipeline.

E.3. Visualization and failure cases

We provide additional visualization of the prediction results (cf. Section 4.5 of the main paper). We consider AVOD with the following point clouds and representations.

- LiDAR
- pseudo-LiDAR (stereo): with PSMNET⋆ [1]
- pseudo-LiDAR (mono): with DORN [3]

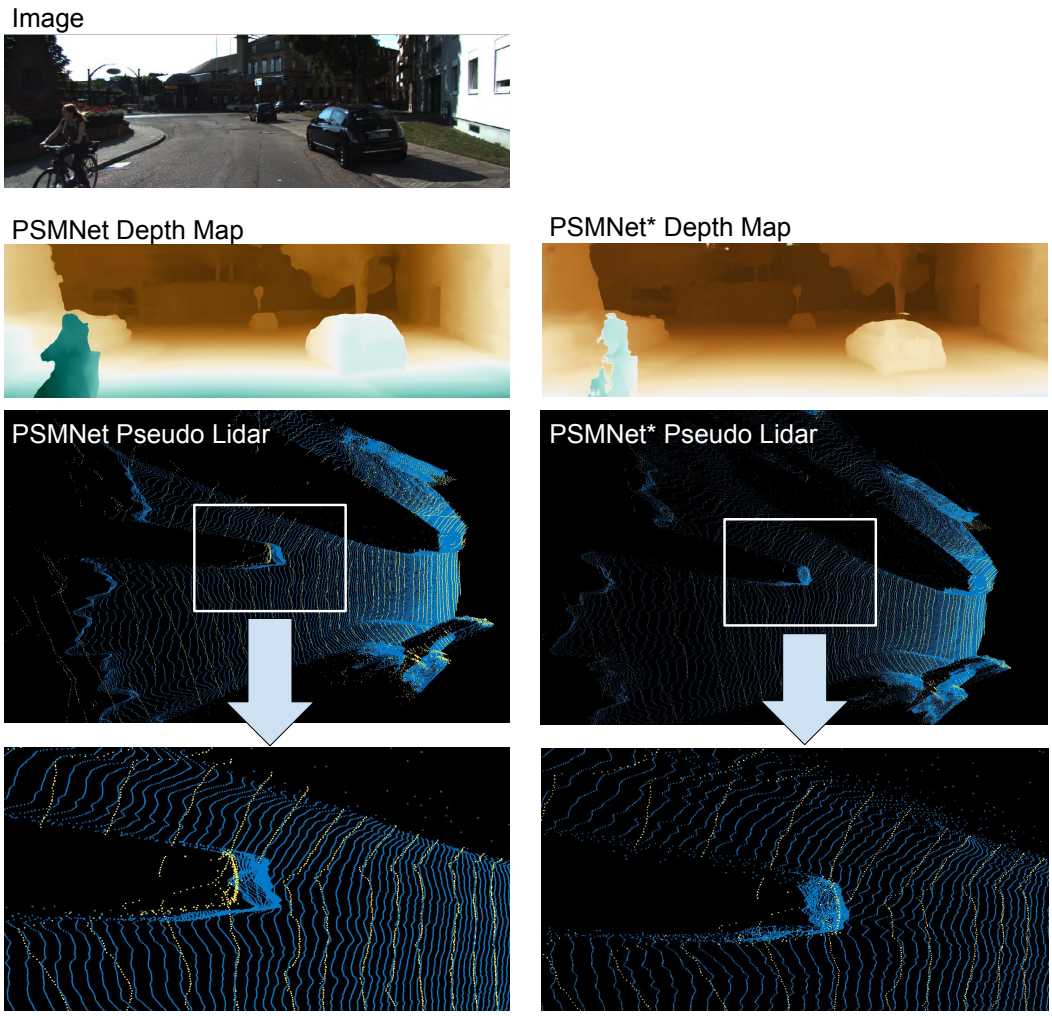


Figure B: **PSMNET vs. PSMNET⋆.** Top: a KITTI street scene. Left column: the depth map and pseudo-LiDAR points (from the bird's-eye view) by PSMNET, together with a zoomed-in region. Right column: the corresponding results by PSMNET⋆. The observer is on the very right side looking to the left. The pseudo-LiDAR points are in blue; LiDAR points are in yellow. The pseudo-LiDAR points by PSMNET have larger deviation at far-away distances. Best viewed in color (zoom in for details).

- frontal-view (stereo): with PSMNET⋆ [1]

We note that, as DORN [3] applies ordinal regression, the predicted monocular depth are discretized.

As shown in Fig. C, both LiDAR and pseudo-LiDAR (stereo or mono) lead to accurate predictions for the nearby objects. However, pseudo-LiDAR detects far-away objects less precisely (**mislocalization**: gray arrows) or even fails to detect them (**missed detection**: yellow arrows) due to in-accurate depth estimates, especially for the monocular depth. For example, pseudo-LiDAR (mono) completely misses the four cars in the middle. On the other hand, the frontal-view (stereo) based approach makes extremely inaccurate predictions, even for nearby objects.

To analyze the failure cases, we show the precision-recall (PR) curves on both 3D object and BEV detection in Fig. D. The pseudo-LiDAR-based detection has a much lower recall compared to the LiDAR-based one, especially for the moderate and hard cases (i.e., far-away or occluded objects). That is, missed detections are one major issue that pseudo-LiDAR-based detection needs to resolve.

We provide another qualitative result for failure cases in Fig. E. The partially occluded car is missed detected by AVOD with pseudo-LiDAR (the yellow arrow) even if it is close to the observer, which likely indicates that stereo disparity approaches suffer from noisy estimation around occlusion boundaries.

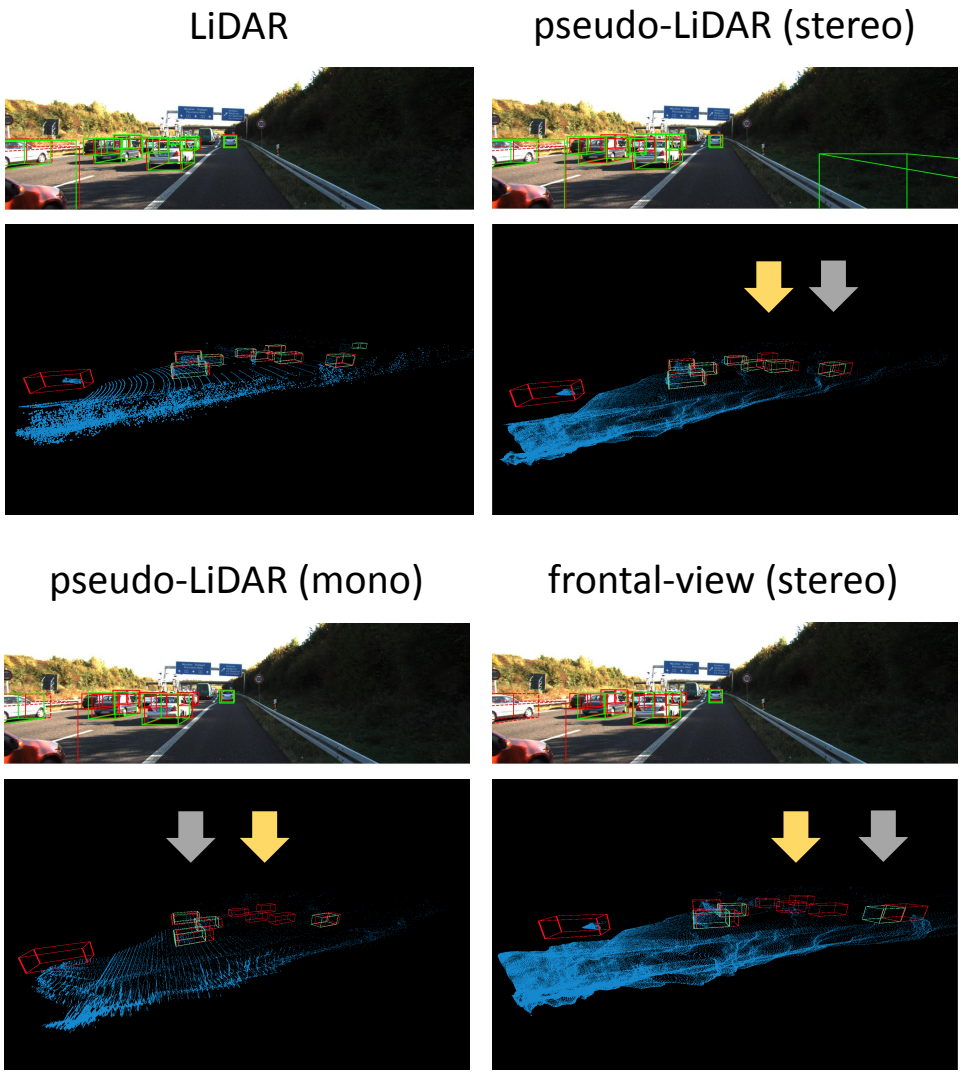


Figure C: **Qualitative comparison and failure cases.** We compare AVOD with LiDAR, pseudo-LiDAR (stereo), pseudo-LiDAR (monocular), and frontal-view (stereo). Ground-truth boxes are in red; predicted boxes in green. The observer in the pseudo-LiDAR plots (bottom row) is on the very left side looking to the right. The **mislocalization** cases are indicated by gray arrows; the **missed detection** cases are indicated by yellow arrows. The frontal-view approach (*bottom-right*) makes extremely inaccurate predictions, even for nearby objects. Best viewed in color.

References

[1] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *CVPR*, 2018. 1, 2, 3

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[3] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2, 3

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

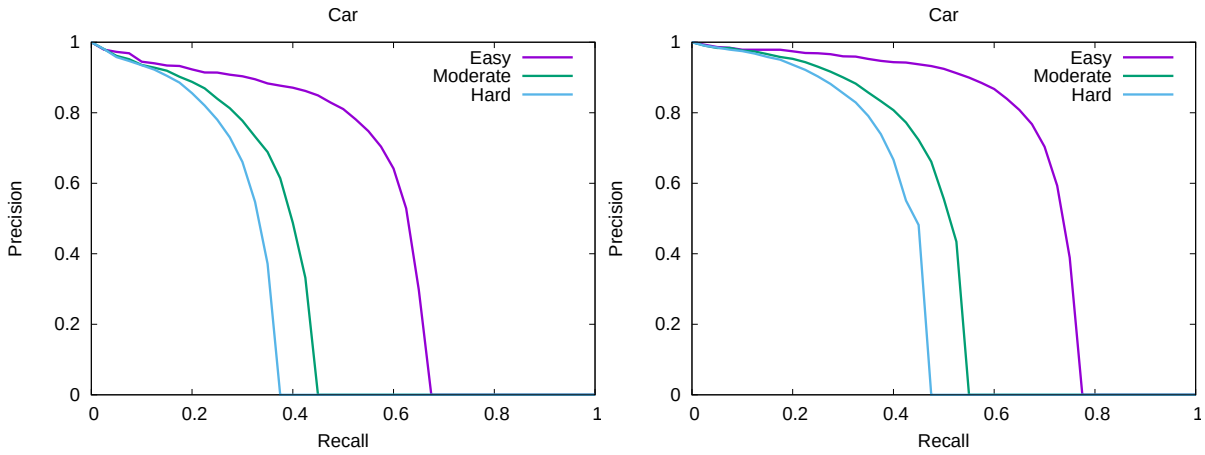
[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[6] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IROS*, 2018. 1

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

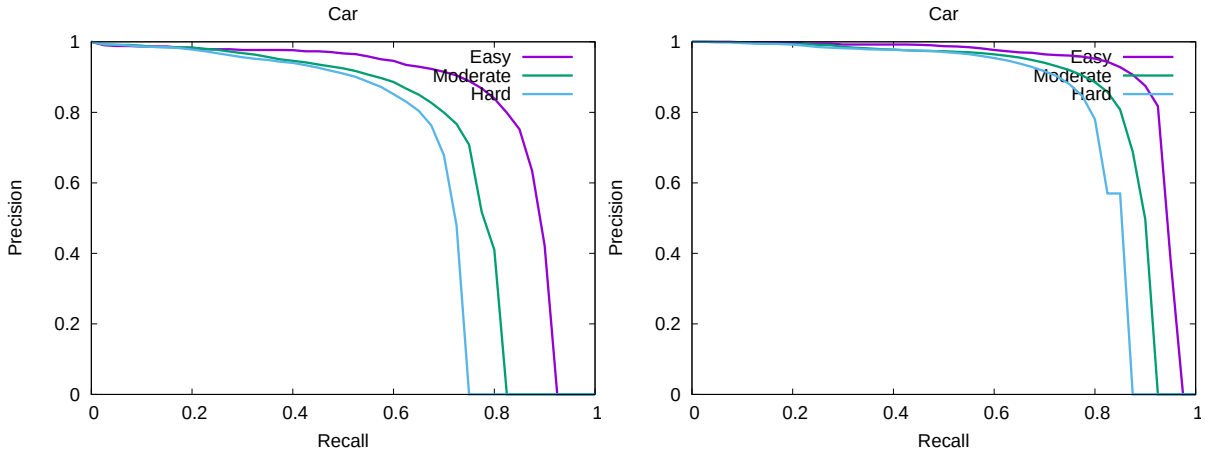
[8] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 1

[9] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum



(a) 3D detection: AVOD + pseudo-LiDAR (stereo)

(b) BEV detection: AVOD + pseudo-LiDAR (stereo)



(c) 3D detection: AVOD + LiDAR

(d) BEV detection: AVOD + LiDAR

Figure D: **Precision-recall curves.** We compare the precision and recall on AVOD using pseudo-LiDAR with PSMNET⋆ (top) and using LiDAR (bottom) on the test set. We obtain the curves from the KITTI website. We show both the 3D detection results (left) and the BEV detection results (right). AVOD using pseudo-LiDAR has a much lower recall, suggesting that missed detections are one of the major issues of pseudo-LiDAR-based detection.

pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 1, 2

[10] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu. Accurate single stage detector using recurrent rolling convolution. In *CVPR*, 2017. 2

[11] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014. 1

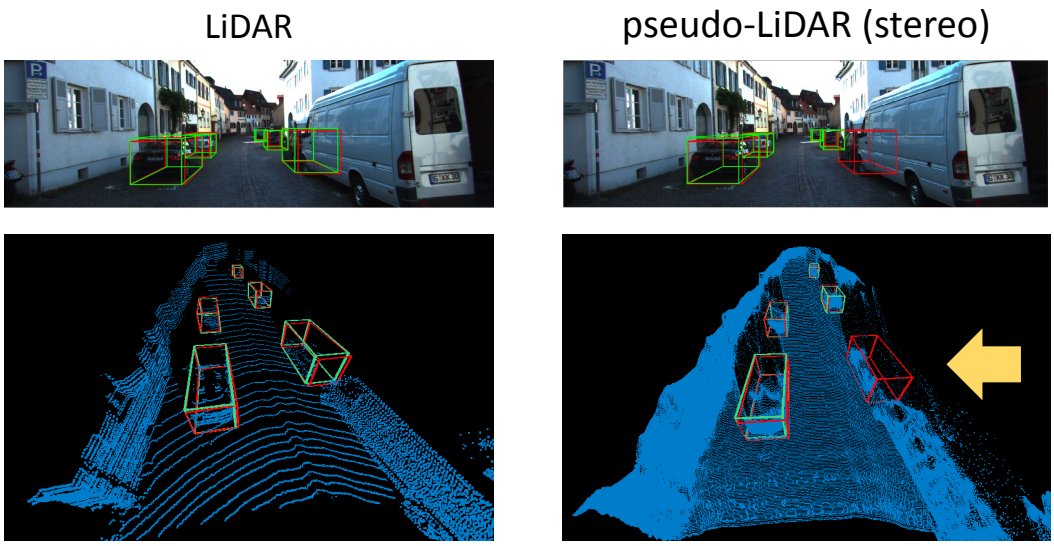


Figure E: **Qualitative comparison and failure cases.** We compare AVOD with LiDAR and pseudo-LiDAR (stereo). Ground-truth boxes are in red; predicted boxes in green. The observer in the pseudo-LiDAR plots (bottom row) is on the bottom side looking to the top. The pseudo-LiDAR-based detection misses the partially occluded car (the yellow arrow), which is a hard case. Best viewed in color.