Decoupled Pseudo-labeling for Semi-Supervised Monocular 3D Object Detection (Appendix)

1. Geometric Relationship in Monocular 3D Object Detection

In contrast to 2D object detection, Monocular 3D Object Detection (M3OD) involves numerous geometric priors between 2D and 3D space, with the primary one being the pinhole model that describes the correspondence between the location of 3D points and 2D points. With the pinhole camera model, the mapping of a 3D point P with $\boldsymbol{c^{\omega}} = (x, y, z)$ location under the LiDAR coordinate system and its corresponding 2D location $\boldsymbol{c^{i}} = (u, v)$ within image can be described as:

$$z[u \;\; v \;\; 1]^T = \mathbf{K} \cdot [\mathbf{R}|\mathbf{T}] \cdot [x \;\; y \;\; z]^T \tag{1}$$

where matrix K is the **camera intrinsic matrix**. z is the depth value at point P, R and T is the rotation and transformation matrix of the **camera extrinsic matrix**. In this equation, the camera extrinsic matrix $[\mathbf{R}|\mathbf{T}]$ is responsible for the transformation between the LiDAR coordinate system and the camera coordinate system. The camera intrinsic matrix K is used to transform the point from the camera coordinate to the image plane.

2. Extended Details of Homography-based Pseudo-label Mining

The complete procedure of the proposed Homography-based Pseudo-label Mining(HPM) algorithm is summarized in Algorithm 1. Several key components of this algorithm are explained below.

Model Prediction. Formally, the outputs of the teacher model F for a unlabeled image I^u contains the 2D and 3D attributes for each predicted object:

$$[Cls, BBox]_{2D}, [Points, Depth, Size, Ori]_{3D} = \boldsymbol{F}(\boldsymbol{I}^u). \tag{2}$$

where Cls is the classification confidence, $BBox$ is the 2D bounding box of the object. $Points$ is the predicted projected points of the 3D bounding box in the image plane. In our main paper, we predict 10 points of the 3D box in total following MonoFlex[12], which includes eight corner points and top and bottom center points. $Depth$ refers to the depth value of the bottom center point. $Size$ represents the 3D size and contains the length, width, and height of the object. Ori represents the orientation of the object.

2D-3D Transformation. Homography-based Pseudo-label Mining involves the geometric transformation of the pseudo-labels between 2D and 3D space. Specifically, we take the four bottom corner points plus the bottom center point as the candidate points for each object to estimate the homography transformation. Specifically, the location of these candidate points in the image plane is directly obtained via Eq.2. To estimate the BEV coordinate of these points, the bottom center point is first transformed from the image plane to the camera coordinate system as:

$$P_{center} = K^{-1} \cdot [zu, zv, z]. \tag{3}$$

Then, we apply the local transformation to the P_c with the orientation and 3D size(length, width, and height) prediction in Eq.2 to get the camera coordinates P_{corner} of the candidate corner points. Such local transformation involves simply translation and rotation. Finally, the inversion of camera extrinsic matrix $[R|T]$ is further applied to obtain their BEV coordinates.

$$B_{\cdot} = [\mathbf{R}|\mathbf{T}]^{-1} \cdot P_{\cdot} \tag{4}$$

Where $P_{\cdot}$ refers to the camera coordinates of bottom corner points and bottom center point, and $B_{\cdot}$ denotes their corresponding BEV coordinates.

Note that the 3D BEV coordinate of these candidate points is not directly transformed by their 2D location in the image plane, instead, they are estimated based on the model's 3D attributes prediction such as depth, orientation, etc. Therefore, the homography matrix solved from these coordinates via DLT[8] is not a trivial solution.

Feasibility of Flat Ground Assumption. We check the feasibility of the flat ground assumption used in the homography-based pseudo-label mining algorithm. Actually, the KITTI dataset does exhibit some micro-unevenness in the ground, which also leads to minor localization errors in ground truth objects when utilizing the homography solved from the ground truth bounding box as shown in Tab.4. However, these errors are minimal when compared to the localization errors in pseudo-labels caused by inaccurate depth. With this obvious gap between the ground

Algorithm 1: $\mathbf{DLT}$ indicates solving the homography matrix by Direct Linear Transform[8]. $\mathbf{ImageCoord}_{.}$ and $\mathbf{BEVCoord}_{.}$ refers to the operation to obtain the coordinates of the point in image plane and BEV plane. $\mathbf{ImageCoord}_{.}$ relies on Eq.2 and $\mathbf{BEVCoord}_{.}$ relies on Eq.3,Eq.4. $\mathbf{'all'}$ indicates all candidate points(5 points each object), $\mathbf{'bc'}$ indicates the bottom center point.

```
Input  :
        $m^0$: Initial set of pseudo-labels generated by uncertainty filtering;
        $\theta_h$: Localization error threshold;
        $p = \{p_1, ..., p_N\}$: $N$ candidate predicted 3D bounding box
Output:
        $m^{3D}$: The final pseudo-labels for 3D attributes

for $t = 1, \cdots, t_{max}$ do
    $\tilde{C}_I^t = \mathbf{ImageCoord_{all}}(m^{t-1})$; // Image coordinates of all pseudo-labels' bottom corner and center points
    $\tilde{C}_B^t = \mathbf{BEVCoord_{all}}(m^{t-1})$; // BEV coordinates of all pseudo-labels' bottom corner and center points
    $\tilde{M} \leftarrow \mathbf{DLT}(\tilde{C}_I^t, \tilde{C}_B^t)$;
    $m^t \leftarrow m^{t-1}$;
    for $j = 1, \cdots, N$ do
        $\hat{c}^b \leftarrow \tilde{M}[\mathbf{ImageCoord_{bc}}(p_j), 1]^T$; // Obtain BEV coordinates of the bottom center points via homography
        transformation
        $\epsilon_j^t \leftarrow ||\mathbf{BEVCoord_{bc}}(p_j) - \hat{c}^b||_2$; // Compute loc error of bottom center point
        if $\epsilon_j^t < \theta_h$ then
            $m^t \leftarrow$ add $b_j$ into $m^t$;
        end if
    end for
    if $m^t = m^{t-1}$ then
        break;
    end if
end for
$m^{3D} \leftarrow m^t$
return $m^{3D}$
```

truth bounding box and the pseudo-labels with inaccurate 3D attribute prediction, our method remains applicable to distinguish the reliable pseudo-labels. But it's worth noting that on a severely uneven road surface, where the homography constraint is substantially violated by the ground truth object, our approach may struggle to distinguish between reliable and unreliable pseudo-labels.

3. Extended Implementation Details

Our experiments are conducted based on the MonoFlex[12] with the official code provided by the authors. For the **KITTI** dataset, we first pre-train the model on the labeled data for 140 epochs with a batch size of 8 following the default setting. After that, we copy the pre-trained model weight into the student and teacher models for end-to-end semi-supervised fine-tuning. For each iteration of semi-supervised fine-tuning, we randomly select 8 labeled images and 8 unlabeled images as the batched data and pad the images to the size of [1280, 384]. We utilize the AdamW optimizer with a learning rate of 3e-4, and weight decay of 1e-5, and fine-tune the model with semi-supervised learning for 20 epochs, in which the learning rate is decayed at the 10th and 15th epochs by a factor of 0.1, respectively. To demonstrate the generality of our approach, we also conduct experiments on the **nuScenes** dataset [1], which is another large-scale autonomous driving dataset. Since [9, 10] are the only M3OD works that provide the results on this benchmark, we choose to conduct experiments with these two base detectors. For the nuScenes dataset, we follow the default setting of FCOS3D[9] and PGD[10] implemented in MMDetection3D. We first pre-train the model on the labeled data for 12 epochs with a batch size of 16 and input size of [1600,900]. We utilize the SGD optimizer with a learning rate of 2e-3 and weight decay of 1e-4. For each iteration of semi-supervised fine-tuning, we randomly select 8 labeled images and 8 unlabeled images as the batched data and conduct fine-tuning for 5 epochs, in which the learning rate is decayed at the 2nd and 4th epochs by a factor of 0.1, respectively. For the experiments on Other M3OD Detectors, given that the pseudo-label mining algorithm based on homograph in DPL relies on key point prediction of the 3D bounding box. For the monocular 3D object detector

[6, 7, 11] without key point prediction, we add the key point prediction branch head to the original head. All experiments are conducted with $8\times$ 32G NVIDIA Tesla V100 GPUs.

4. More Experiment Results

Effect of Diversity of Unlabeled Data. We analyze the impact on the SSM3OD performance of the distribution of the unlabeled data. The KITTI raw data were collected from five diverse scenes: city, residential, road, campus, and person, as depicted in the right of Fig.1. Analyzing the object class distribution in each scene revealed significant differences. For example, in residential and road scenes, the car object dominates, with few pedestrians and cyclists. Conversely, campus and personal scenes mainly consist of pedestrians. To investigate the effect of unlabeled data diversity, the images were roughly divided into two groups: car-oriented, and person-oriented according to their object class distribution. We randomly chose 5K images for each group as the unlabeled data. We further constructed a more comprehensive unlabeled data set by combining the images randomly selected from both two groups, with 2.5K images selected for each group. The results are reported in Tab.1. It clearly shows that the diversity of these classes affects performance. Specifically, car-dominated unlabeled images boost the performance of the car category, while resulting in a slight performance decrease for the pedestrian and cyclist category. There is an opposite trend when training with person-dominated unlabeled images, in which the performance of pedestrian and cyclist categories are improved and no obvious performance gain for the car category is observed. The main reason behind these results is the confirmation bias caused by the class imbalance. By contrast, the unlabeled data combined images from both groups, containing rich objects, improve the performance of all three object categories. These results underscore the importance of unlabeled data diversity in semi-supervised learning for M3OD.

Performance on Large Scale Dataset. The nuScenes Dataset is a large dataset for multi-view 3D object detection (MVOD), and some M3OD methods [9, 10] can be extended to achieve MVOD by conducting monocular detection in every single view and then fusing the multi-view detection results. To demonstrate the generality of our method, we conducted further experiments on this large-scale dataset based on FCOS3D and PGD with the official codes from MMDetection3D. Note MVC-MonoDet [4] is the only SSM3OD work that reports the results on the nuScenes dataset, but they only present some partial metrics. Our results, detailed in Tab.2, showcase substantial performance improvements through our proposed pseudo-labeling method. Specifically, we achieve gains of **3.1** in mAP and **2.2** in NDS for FCOS3D, **1.8** in mAP, and **1.2** in NDS for PGD. Our method also outperforms MVC-MonoDet in both mAP and mATE metrics. These results verify the efficacy of our method, demonstrating its potential for extending to multi-view 3D object detection and generalization. Note that MVOD focuses on feature intersection between different views or temporal frames which is beyond the scope of this article. Therefore, we *not aim to surpass the state-of-the-art methods [2, 3, 5] on this benchmark* , and instead just to show the generalization ability of our method.

Table 1. The effect of the diversity of unlabeled data. For each group, we randomly select 5K images as the corresponding unlabeled data, respectively. We further construct a comprehensive unlabeled data set by combining both images randomly selected from Car-dominated and person-dominated groups, with 2.5K images randomly chosen for each group.

Unlabeled Data	Val, $AP_{3D}\|R_{40}$								
	Car			Pedestrian			Cyclist		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Sup-baseline	22.80	17.51	14.90	7.30	5.53	4.24	4.67	2.23	1.93
Car-dominated	**24.54**	18.44	15.68	6.91	5.43	4.51	3.52	2.04	1.65
Person-dominated	22.76	17.23	14.71	8.23	6.66	5.02	4.87	2.32	2.11
Combined	24.32	**18.56**	**16.12**	**8.45**	**6.72**	**5.07**	**5.35**	**2.89**	**2.43**

Ablation of Threshold. We ablate the threshold of uncertainty threshold θ_u, location error threshold θ_h in Tab.3. The best results achieve with $\theta_u = 0.10$ and $\theta_h = 2.0$.

Performance on Pedestrian and Cyclist Categoriy. We also report the detection performance on the Pedestrian and Cyclist categories on the KITTI test set in Tab.5, where our method also provides a significant boost in detection performance for these categories with relatively few instances.

Detection Results Visualization. We visualize the detection results of our method compared to the supervised baseline method on the KITTI validation set in Fig.2. It clearly shows that our method not only detects objects more accurately, as observed in 1st, 2nd, and 3rd images but also exhibits higher prediction recall, as presented in 4th and 5th images. These results once again demonstrate the superiority of our method.

5. Limitations

Our method significantly improves the performance of monocular 3D detection methods with only image input. Compared with other 3D object detection methods, for example, LiDAR-based method, BEV-based method, etc, monocular 3D object exhibits a great advantage in the practical application. With single-camera setups, it is more cost-effective and adaptable to numerous practical scenarios such as robotics, autonomous driving, and mobile augmented reality. However, due to the fundamental difficulty in estimating depth from a single RGB image, current performance still lags behind some methods using extra inputs (such as LiDAR). This limitation motivates us to ex-

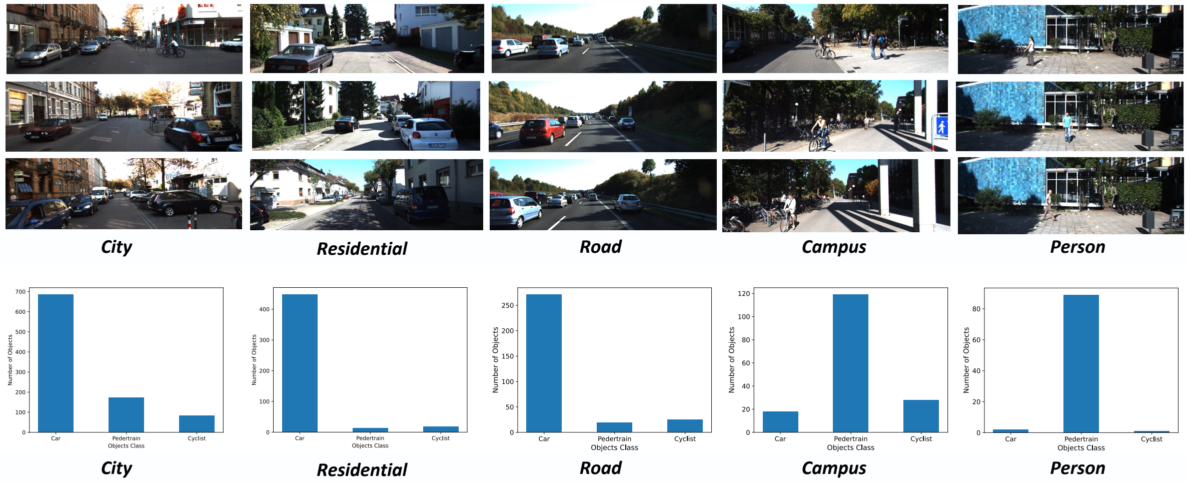

Figure 1. KITTI raw data spans five scene types. Bottom: Object class distribution across scenes. Accordingly, We divide the unlabeled data into two groups: (1) Car-dominated (city, residential, road), and (2) Person-dominated (campus, person).

Table 2. The performance comparison in the **nuScenes validation set**. Note that the MVC-MonoDet only provides the mAP and mATE metrics in this dataset.

Methods	Extra Data	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	NDS ↑
		(%)	(m)	(1-iou)	(rad)	(m/s)	(1-acc)	%
MVC-MonoDet	Unlabeled	0.349	0.640	-	-	-	-	-
FCOS3D	None	0.321	0.754	0.260	0.486	1.332	0.157	0.394
DPL$_{FCOS3D}$	Unlabeled	0.352	0.633	**0.248**	0.423	1.264	**0.143**	0.416
PGD	None	0.358	0.667	0.264	0.434	1.276	0.176	0.425
DPL$_{PGD}$	Unlabeled	**0.376**	**0.577**	0.250	**0.412**	**1.258**	0.161	**0.437**

θ_u	θ_h	Val, $AP_{3D}\|R_{40}$		
		Easy	Mod	Hard
0.10	1.0	25.68	19.39	17.06
0.20	2.0	24.94	19.15	16.82
0.10	2.0	**26.51**	**19.84**	**17.13**

Table 3. Ablation of the threshold θ_u and θ_h.

Labels	Loc Error(m)
GTs	0.91
PLs	2.29

Table 4. The average localization errors of pseudo labels(PL) and ground truth(GT) labels.

Table 5. Performance comparison on the **KITTI test** set of the Pedestrian and Cyclist category.

Methods	Test, $AP_{3D}\|R_{40}$					
	Pedestrian			Cyclist		
	Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN	4.92	3.48	2.94	0.94	0.65	0.47
MonoPair	10.02	6.68	5.53	3.79	2.12	1.83
MonoFlex †	9.02	6.13	5.14	2.36	1.44	1.07
DPL$_{FLEX}$	**11.66**	**7.52**	**6.16**	**8.41**	**4.51**	**3.59**

plore the use of unlabeled data from other complementary sensor modalities such as LiDAR point clouds and stereo images. These unlabeled data contain more reliable object depth information, which can greatly ease the difficulty of accurately detecting 3D objects in real-world scenarios. We leave this exploration for our future work.

References

[1] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[2] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[3] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 3

[4] Qing Lian, Yanbo Xu, Weilong Yao, Yingcong Chen, and Tong Zhang. Semi-supervised monocular 3d object detection by multi-view consistency. In *Computer Vision–ECCV*

2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII, pages 715–731. Springer, 2022. 3

[5] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781. IEEE, 2023. 3

[6] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 3

[7] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 3

[8] Robert Shapiro. Direct linear transformation method for three-dimensional cinematography. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, 49(2):197–205, 1978. 1, 2

[9] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 2, 3

[10] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2, 3

[11] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022. 3

[12] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 1, 2

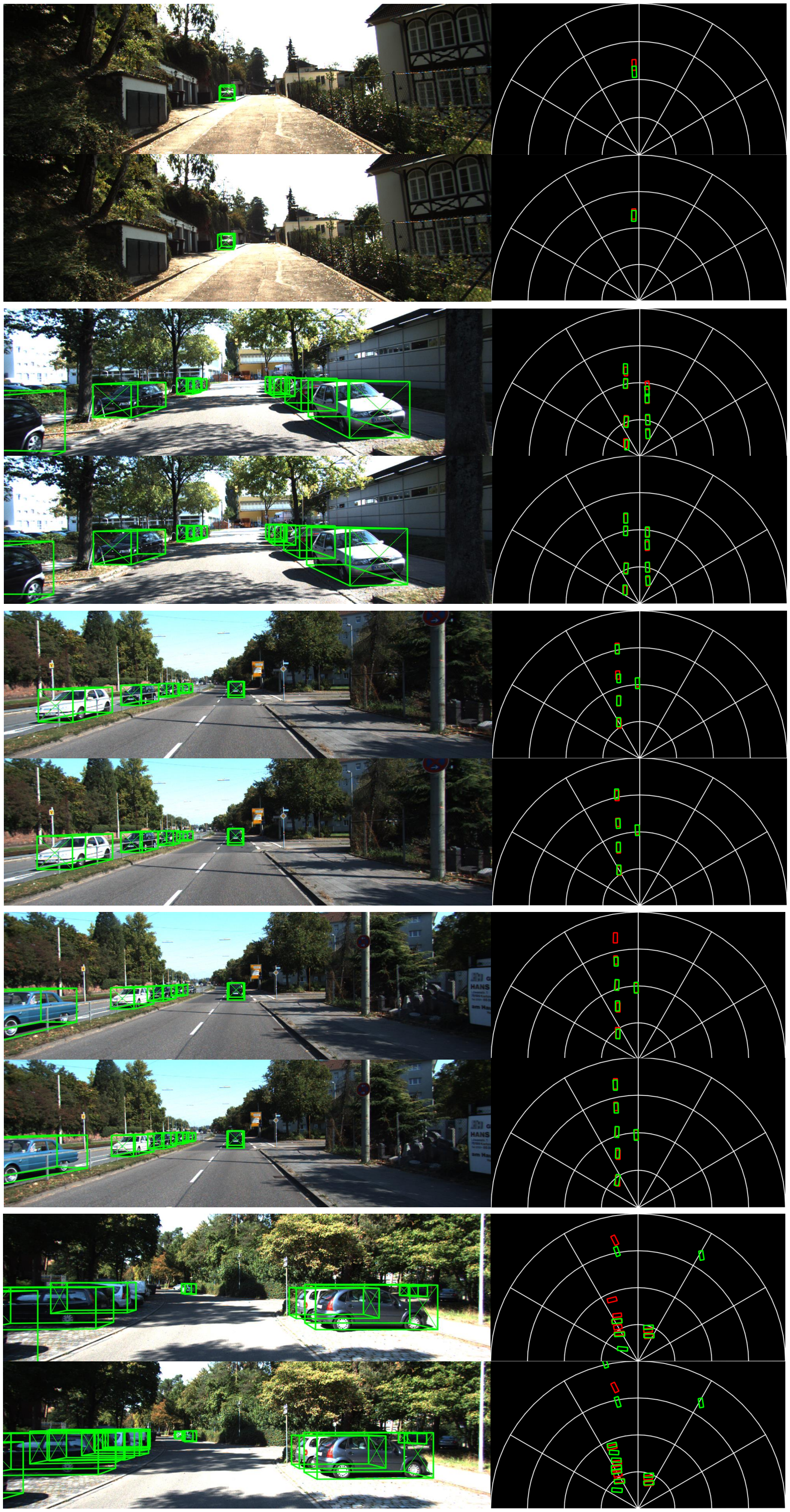

Figure 2. Visualization of the detection results on the KITTI validation set. For each image, the first and second row is the detection results of the supervised baseline and our method, respectively. The box in red and green on the BEV plane are the ground truth box and detection bounding box, respectively.