Towards Robust 3D Object Detection with LiDAR and 4D Radar Fusion in Various Weather Conditions

Supplementary Material

This supplementary material provides additional implementation details, an explanation of design choice on K_l nearest neighbor voxel features, additional quatitative SoTA comparison, feature activation analysis of our model and additional qualitative results on various weather conditions. In particular, the following contents are included in the supplementary material:

- Implementation details of the proposed method.
- Explanation of design choice on K_l and r_l.
- Additional SoTA comparison.
- Analysis on feature activation.
- Additional qualitative results.

1. Additional Implementation Details

Our framework is implemented with PyTorch and trained with an A6000 GPU. We first use batch size 16 and SGD optimizer for the image-based weather classification network with lr=1e-3 and β_1=0.9. Then, we use batch size 4 and Adam optimizer with lr=1e-3, β_1=0.9, β_2=0.999, weight decay 0.01 and the cosine annealing scheduler with the minimum learning rate 1e-4 while training the entire network. The image-based weather classification network comprises three convolutional layers, sequentially downsizing the spatial dimensions by half while increasing the channel dimensions to [16, 32, 64]. Batch normalization layers are inserted every after convolutional layers. Then, the feature is flattened and passes through a linear layer, which adjusts the channel dimension. The final output is calculated after a GAP layer and a linear layer, and trained with cross-entropy loss. The "repeat" function in WRGNet is designed to replicate the image features to match the dimensions of voxel features. For 4D radar visualization, normalization along the z-axis is initially performed, followed by a logarithmic transformation (base 10), and the resultant data is visualized using a jet colormap. The proposed method currently runs around 5fps, with 0.17s spent on network inference (feature extraction: 25ms, 3D-LRF: 128ms, WRGNet: 1.5ms, BEV encoder: 15ms) and 0.03s on NMS. The current algorithm is not optimized for inference speed. Using TensorRT or adopting a detection head without NMS will enable real-time operation.

2. Design choice of K_l and r_l

We examine the effect of parameters K_l and r_l in the 3D-LRF module. The parameter r_l is set to a proper small value that must contain K_l radar neighbor voxels and does not influence the performance. Therefore, we varied the parameter K_l to investigate its impact on performance, and the results are summarized in Table 1. Initially, we conducted experiments with fixed values of K_l=32 and r_l=4, which served as our baseline. Then, we conducted experiments to assess the impacts of adjusting K_l for each layer in our model. When we set K_l and r_l to smaller values, $\lfloor \frac{32}{2^{l-1}} \rfloor$ and $\lfloor \frac{8}{2^{l-1}} \rfloor$ where $l \in \{1, 2, 3\}$, the overall performance has been slightly degraded due to the limited information exchange between LiDAR and radar modalities. When we set K_l and r_l to larger values, $\lfloor \frac{128}{2^{l-1}} \rfloor$ and $\lfloor \frac{8}{2^{l-1}} \rfloor$, the performance has been slightly increased. And our final setting, K_l=$\lfloor \frac{64}{2^{l-1}} \rfloor$ and r_l=$\lfloor \frac{8}{2^{l-1}} \rfloor$, shows the best performance. The proposed 3D-LRF module exhibits robustness to parameter changes, and furthermore, the proposed strategy of adjusting the K_l adequately according to the layer and feature size aids the overall performance.

Table 1. Effect of K_l and r_l in our 3D-LRF module. Best in **bold**, second in underline.

Parameters		IoU=0.3		IoU=0.5	
K_l	r_l	AP_{3D} (↑)	AP_{BEV} (↑)	AP_{3D} (↑)	AP_{BEV} (↑)
32	4	73.4	83.5	38.2	**72.8**
$\lfloor \frac{32}{2^{l-1}} \rfloor$	$\lfloor \frac{8}{2^{l-1}} \rfloor$	73.2	82.7	37.1	71.8
$\lfloor \frac{128}{2^{l-1}} \rfloor$	$\lfloor \frac{8}{2^{l-1}} \rfloor$	<u>73.7</u>	<u>83.8</u>	<u>38.3</u>	72.1
$\lfloor \frac{64}{2^{l-1}} \rfloor$	$\lfloor \frac{8}{2^{l-1}} \rfloor$	**74.7**	**84.6**	**38.6**	<u>72.7</u>

3. Additional Quatitative Comparison

We implemented other SoTA methods for comprehensive comparison. Recent LiDAR-based VoxelNext [1], LiDAR-/image-based PointAugmenting [5] (point-based fusion) and BEVFusion [3] (BEV-based fusion), and BEVFusion* which is modified to incorporate 4D radar as an additional input are adopted for comparison. As shown in Table 2, ours achieves the best accuracy under all metrics in K-Radar dataset. This demonstrates that our method is more robust than InterFusion [6] and BEVFusion* that process 4D radar identically with LiDAR, and [3, 5] that directly takes images as input. PointAug. [1] achieves second-best AP_{3D} with the aid of virtual points in normal conditions (*e.g.*, 75.0 AP_{3D} under IoU=0.3), however, it does not exhibit

Table 2. Additional SoTA comparison on K-Radar dataset (L: LiDAR, C: camera, R: 4D radar). Best in **bold**, second in underline.

Methods	Mod.	IoU=0.3		IoU=0.5	
		AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}
VoxelNext [1]	L	68.8	80.8	33.7	71.6
PointAug. [5]	LC	73.0	75.3	37.6	68.9
BEVFusion [3]	LC	66.2	78.0	29.9	68.9
BEVFusion*	LCR	70.4	81.6	30.5	72.8
Ours	LR	**74.8**	**84.0**	**45.2**	**73.6**

high performance across various weather conditions. BEVFusion* achieves second-best AP_{BEV} with the aid of 4D radar, however, it does not perform effective multi-modal fusion in 3D as our method, leading to a lower AP_{3D}. While it would be beneficial to compare with more 3D object detection methods utilizing both 4D radar and LiDAR, we were limited to InterFusion as there are no other existing alternatives.

4. Feature Activation Analysis

The feature activation of proposed model is visualized in Fig. 1. We can see that LiDAR activates in areas resembling a car with precise location in (b) and 4D radar roughly activates in areas where any objects are present under all weather in (c). 4D radar helps locally enhance or suppress LiDAR as shown in (d). Weather-conditioned image feature and radar feature are gated to compute (e) and modulate the flow from (d) to (b). The visualized feature activation validates that our framework effectively integrates the characteristics of each modality. Moreover, under normal conditions, as in case I, G_1 from WRGNet does not exhibit significant activation for the radar flow modulation, because LiDAR alone is capable of producing substantial results. Under adverse conditions, as in case II, activation in G_1 facilitates the flow from radar to LiDAR in locations where target exists. This demonstrates that WRGNet is a crucial element enabling robust multi-modal fusion across various weather conditions. Overall, the role of LiDAR in our framework is to detect objects from precise 3D geometry and shape information, while radar selects object candidates based on robust measurements to identify objects missed by LiDAR. The camera determines weather conditions through semantic information and adjusts the flow for LiDAR and 4D radar fusion.

5. Additional Qualitative Results

We present the qualitative results of our and competing models, RTNH [4], RTNH*, PointPillars [2] and InterFusion [6], to provide additional insights into the qualitative performance comparison. Figs. 2 to 8 show the results under normal, overcast, fog, rain, sleet, light snow and heavy snow, respectively. We qualitatively demonstrate that our model accurately detects objects in 3D across various (normal and adverse) weather conditions, surpassing competing models.

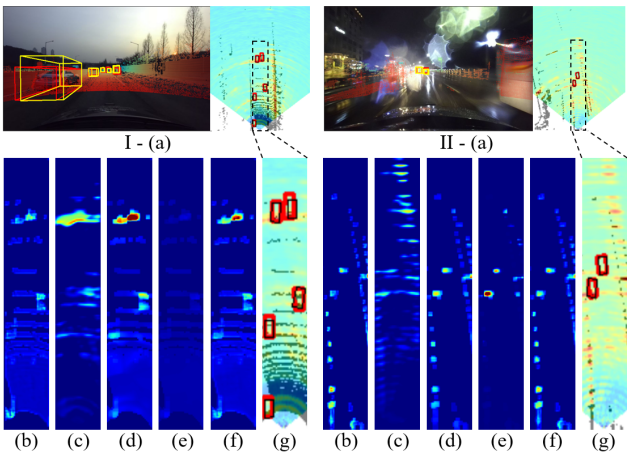

Figure 1. (a) Detection results and feature activation of (b) L_1 (c) R_1 (d) F_1 (e) G_1 (f) $\hat{L_1}$ under normal (I) and rain (II) conditions.

References

[1] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *CVPR*, 2023. 1, 2

[2] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12689–12697, 2018. 2, 3, 4, 5, 6, 7, 8, 9

[3] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. 1, 2

[4] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3, 4, 5, 6, 7, 8, 9

[5] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021. 1, 2

[6] Li Wang, Xinyu Zhang, Baowei Xv, Jinzhao Zhang, Rong Fu, Xiaoyu Wang, Lei Zhu, Haibing Ren, Pingping Lu, Jun Li, and Huaping Liu. Interfusion: Interaction-based 4d radar and lidar fusion for 3d object detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12247–12253, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

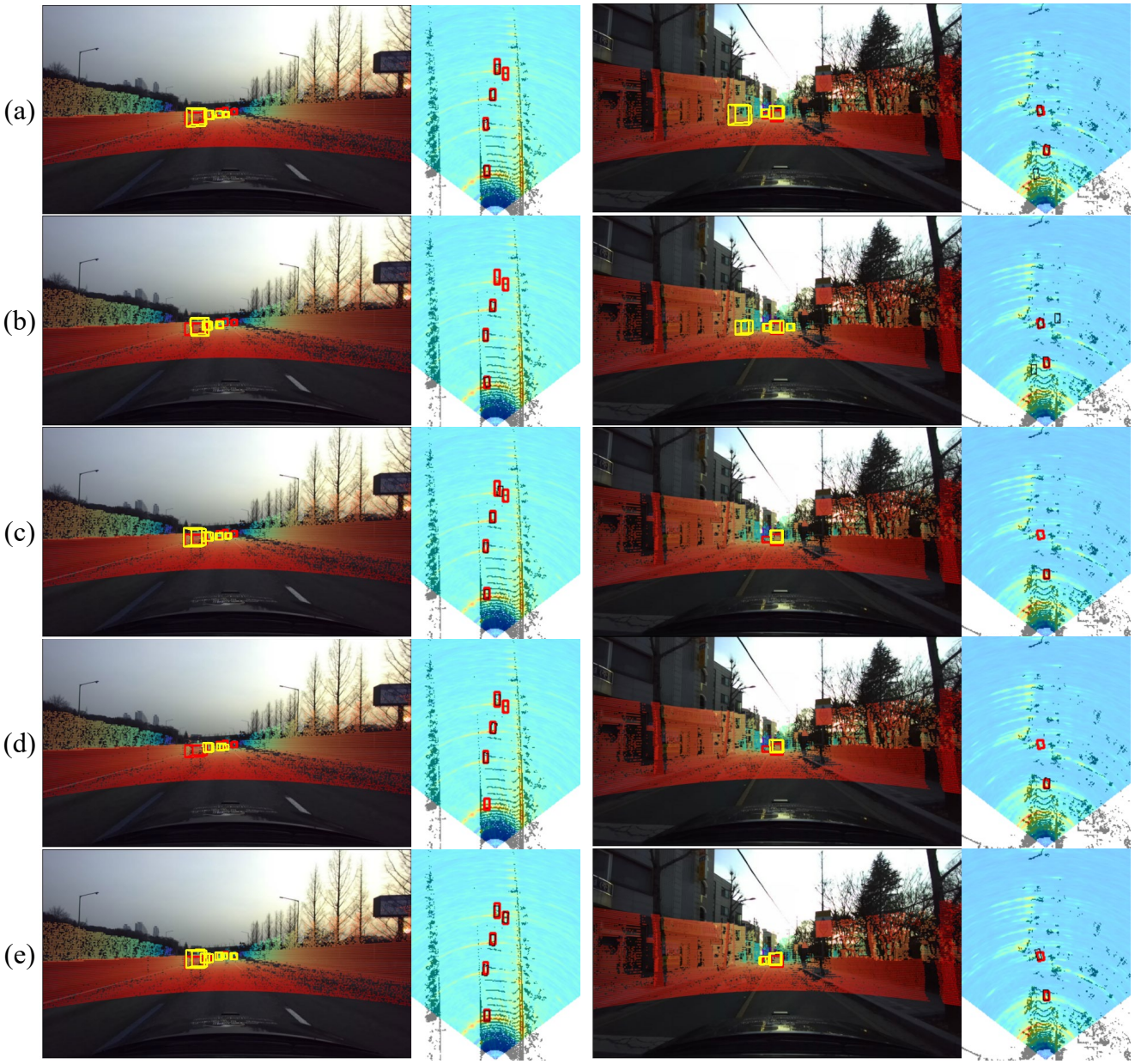

Figure 2. More visual results of 3D object detection in range view and bird-eye-view under "normal" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

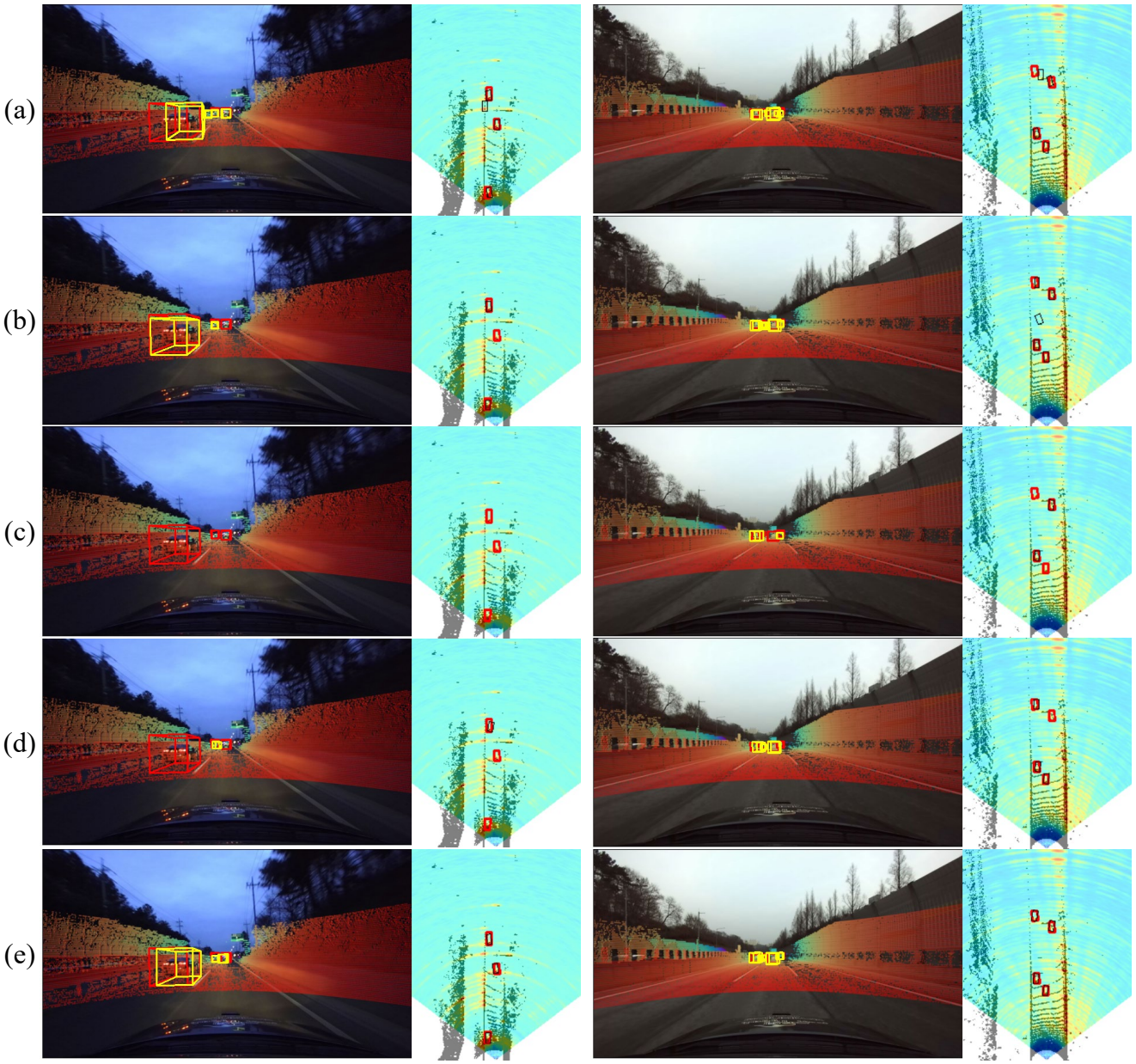

Figure 3. More visual results of 3D object detection in range view and bird-eye-view under "overcast" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

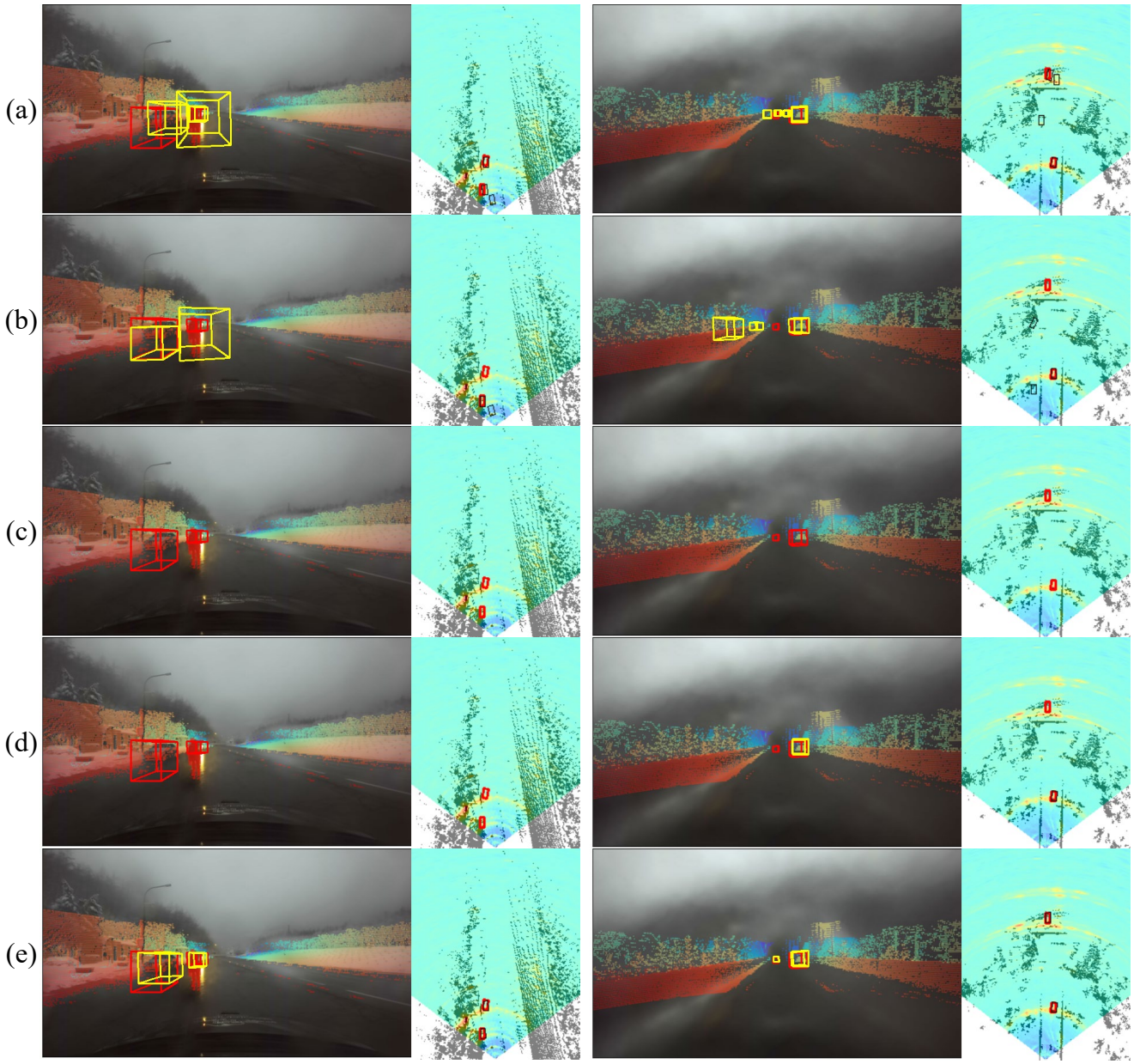

Figure 4. More visual results of 3D object detection in range view and bird-eye-view under "fog" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

Figure 5. More visual results of 3D object detection in range view and bird-eye-view under "rain" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

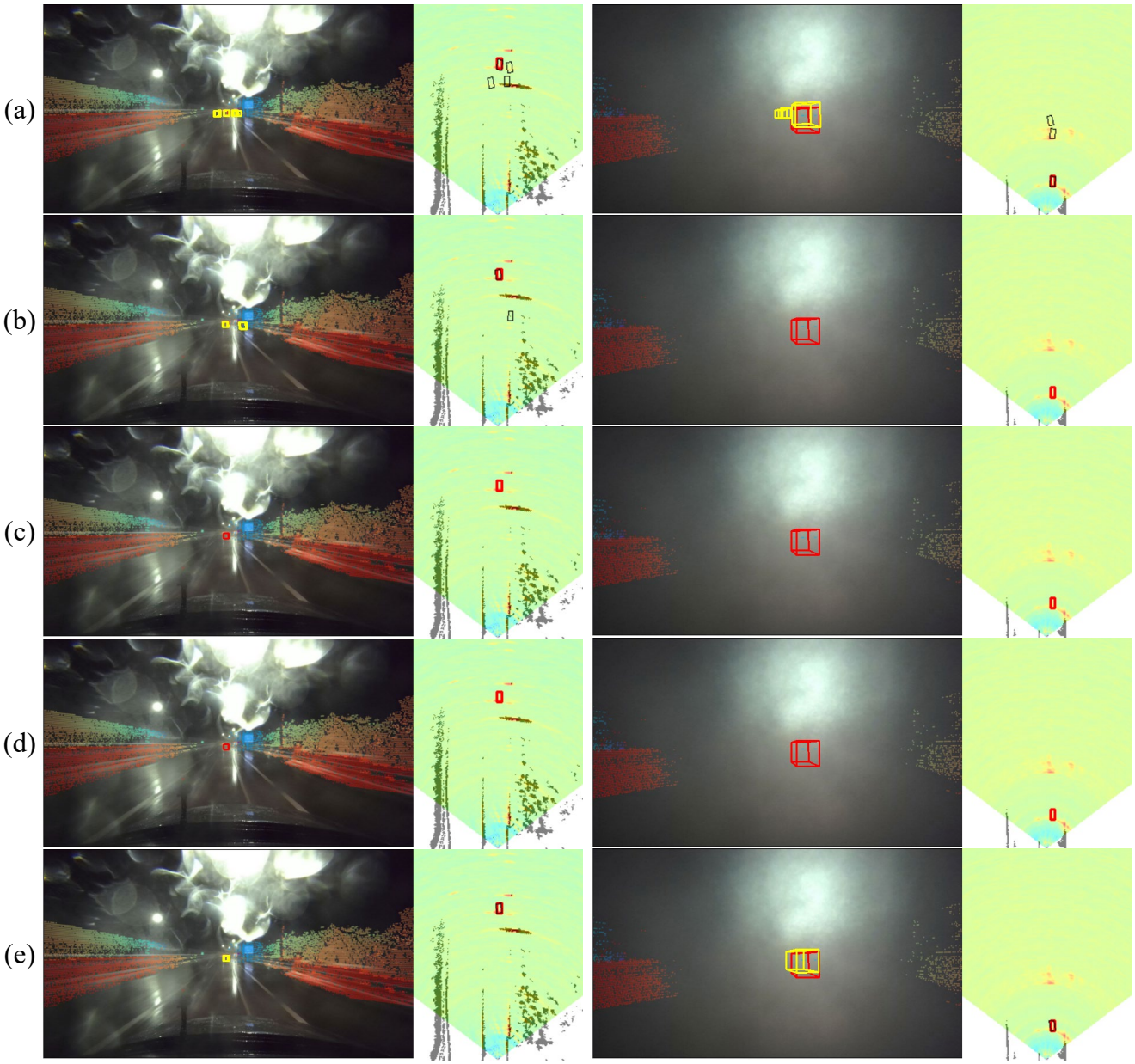

Figure 6. More visual results of 3D object detection in range view and bird-eye-view under "sleet" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

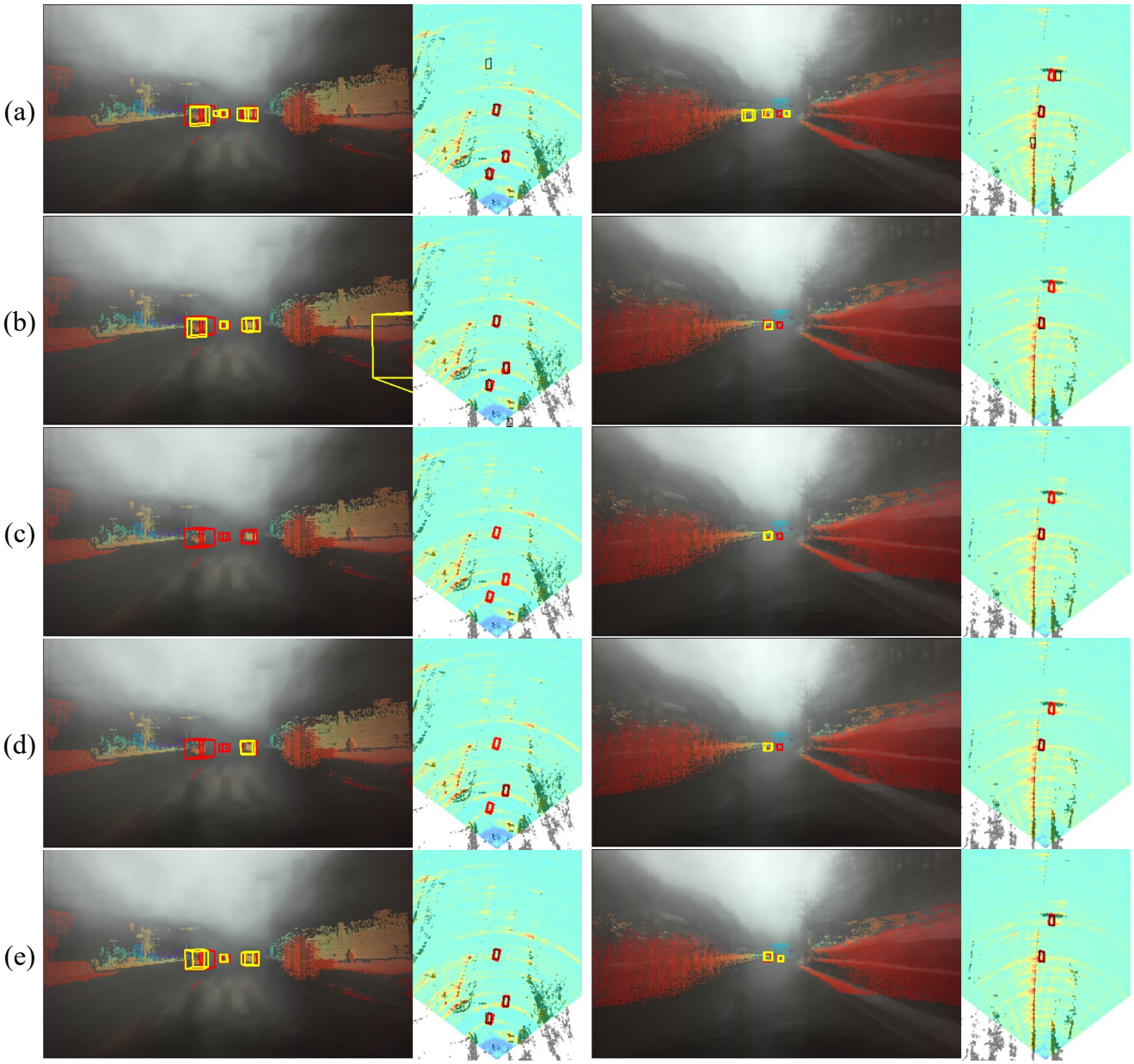

Figure 7. More visual results of 3D object detection in range view and bird-eye-view under "light snow" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.

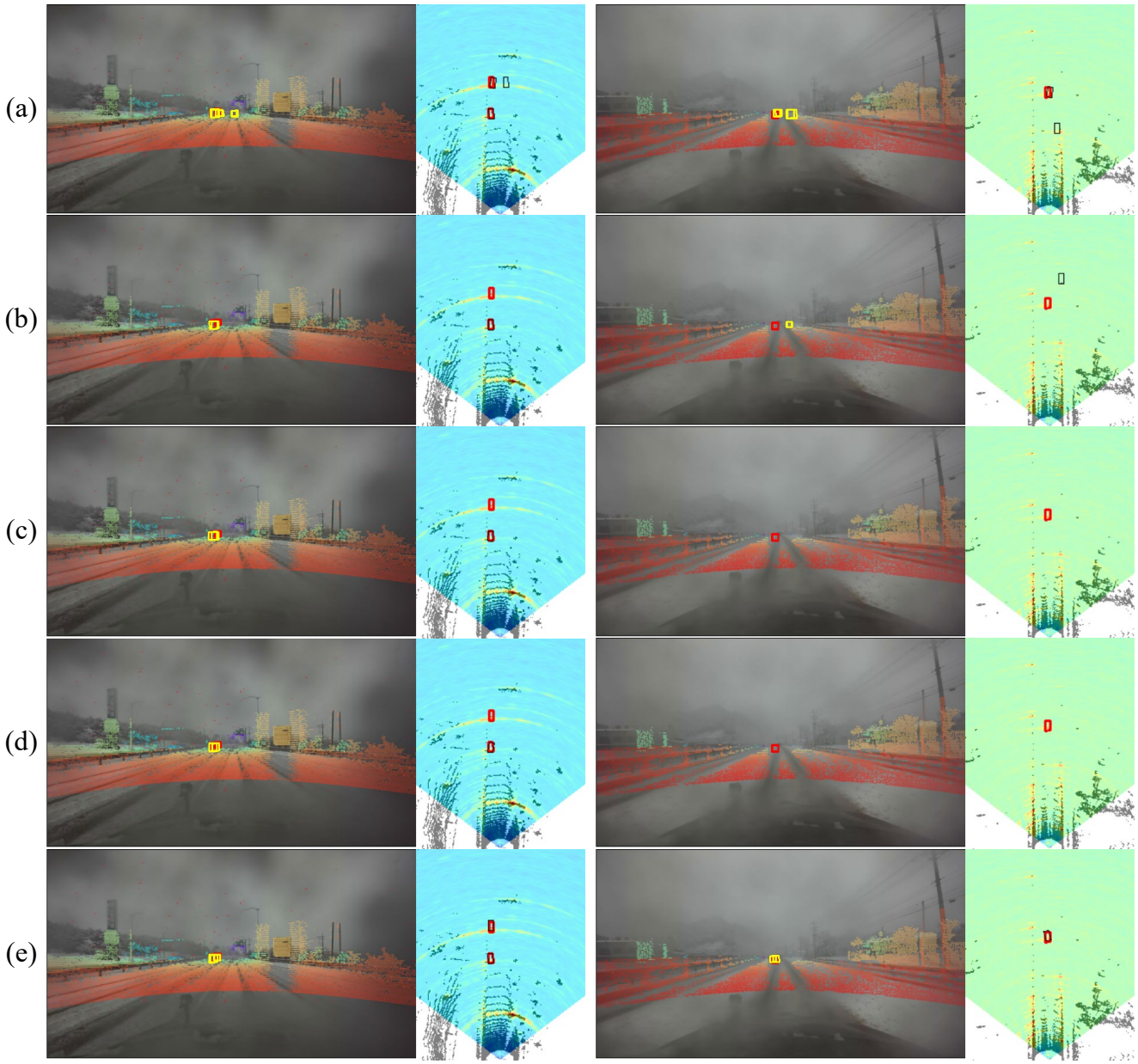

Figure 8. More visual results of 3D object detection in range view and bird-eye-view under "heavy snow" condition. The results in the range view show the image and projected LiDAR with red GT boxes and yellow predicted boxes. The results in bird-eye view show top-view LiDAR and 4D radar heatmap with red GT boxes and black predicted boxes. Each column means the 3D object detection model: (a) RTNH [4], (b) RTNH*, (c) PointPillars [2], (d) InterFusion [6], (e) ours. Best viewed when zoomed in with colors.