LiDAR-based All-weather 3D Object Detection via Prompting and Distilling 4D Radar

–Supplementary materials–

Yujeong Chae, Hyeonseong Kim, Changgyoon Oh, Minseok Kim, and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST
{yujeong, brian617, changgyoon, alstjrx1x1, kjyoon}@kaist.ac.kr

1 Additional Implementation Details

RTNH [9] and RTNH* (baseline) employ the same network architecture, but differ in their input domains (4D radar / LiDAR). Their embedding layer of input domain is composed of a 3D sparse convolution layer [2], and each voxel encoder contains three 3D sparse convolution layers followed by batch normalization [3] and ReLU activation. The BEV encoder for each voxel layer's output consists of a 3D sparse convolution layer followed by batch normalization, which handles sparse voxels. It is then followed by a transposed 2D convolution layer with batch normalization and ReLU activation, which is used to process the densified voxels and adjust their size to match the BEV feature size of all layers. The detection head comprises a classification and regression head, each composed of one convolution block. During the training of the baseline network, we utilize a batch size of 4 and the Adam optimizer [4] with a learning rate (lr) of 1e-3, β_1 of 0.9, and β_2 of 0.999. Additionally, we apply a weight decay of 0.01 and employ the cosine annealing scheduler [8] with a minimum learning rate of 1e-4. The baseline network is trained until convergence using the loss described in Eq. 9 of the main paper. During the training of the teacher network, all parameters of the baseline are frozen, and only the adaptation layers involved in prompt learning are trained. We used the same batch size, optimizer settings, and loss terms and trained until convergence. During the training of the student network, all parameters of the teacher network are frozen, and only the student's embedding, backbone, and head are trained. The teacher and student models do not share any parameters. We used the same batch size and optimizer settings as before and trained until convergence with the loss described in Eq. 10 of the main paper. We implement our framework using PyTorch and train it on an A6000 GPU. To visualize the 4D radar data, we initially normalize along the z-axis, then apply a logarithmic transformation with a base of 10. Finally, the transformed data is visualized using a jet colormap.

Table 1: Quantitative results of LiDAR and 4D radar-based 3D object detection methods on K-Radar dataset [9] (L: LiDAR, 4DR: 4D radar) at IoU=0.5. We present the detailed performance for each weather condition. Best in **bold**, second in underline.

Methods	Modality	Metric	Total	Normal	Overcast	Fog	Rain	Sleet	Lightsnow	Heavysnow
RTNH [9]	4DR	AP_{BEV}	44.4	44.6	61.1	51.6	47.1	28.8	53.8	42.5
		AP_{3D}	14.1	15.5	27.5	21.7	9.21	6.17	28.4	10.6
RTNH* [9]	L	AP_{BEV}	65.8	69.4	86.8	82.1	72.9	42.3	78.0	42.3
		AP_{3D}	37.2	39.8	41.4	58.5	28.2	24.5	50.4	21.2
PointPillars [6]	L	AP_{BEV}	47.8	54.3	60.6	43.5	60.1	18.5	43.3	28.6
		AP_{3D}	21.5	26.4	32.4	29.6	25.7	5.97	28.3	7.50
VoxelNext [1]	L	AP_{BEV}	70.5	70.6	87.2	80.3	79.0	37.4	82.1	48.0
		AP_{3D}	32.7	29.8	35.9	34.7	33.8	9.98	54.9	20.4
InterFusion [10]	4DR+L	AP_{BEV}	59.3	56.4	83.4	68.7	65.3	27.5	77.6	51.7
		AP_{3D}	22.4	20.4	30.8	33.6	20.8	10.6	53.6	19.8
BEVFusion* [7]	4DR+L	AP_{BEV}	71.6	70.7	88.8	80.7	**82.1**	42.6	**84.0**	**52.2**
		AP_{3D}	29.0	30.1	38.7	40.1	32.2	9.20	50.9	**25.3**
Robo3D [5]	L	AP_{BEV}	(55.6)	-	82.8	54.7	65.5	24.1	66.6	41.0
		AP_{3D}	(21.6)	-	51.8	20.9	23.3	3.06	29.9	20.3
Ours-T	4DR+L	AP_{BEV}	71.3	72.1	87.3	83.9	75.2	**45.1**	79.1	48.1
		AP_{3D}	40.4	40.3	43.3	63.4	31.6	**27.9**	51.6	24.9
Ours-S	L	AP_{BEV}	**72.4**	**74.0**	**89.3**	**85.7**	74.6	39.6	79.9	46.3
		AP_{3D}	**43.3**	**47.4**	**55.0**	**68.6**	**36.3**	23.2	**56.8**	25.2

2 Additional Quantitative Results

We provide the additional quantitative comparison on K-Radar dataset at IoU= 0.5 in Table 1. When IoU=0.5, similar to IoU=0.3, Ours-T and Ours-S demonstrate noticeable improvements in performance across various weather conditions through an effective prompt-based fusion and distillation methods. The total results of Robo3D (shown in parentheses) represent the average performance solely under adverse weather conditions.

3 Additional Model Analyses

Component Analysis of Prompt Learning. We conducted the experiments three times and reported mean and standard deviation (std) in Table 3 in the main paper. By running the experiments multiple times, it was concretely demonstrated that both self-calibration (SC) and global aggregation (GA) individually improve performance compared to (a). However, (b) and (c) show higher std. This is because without SC, the features from the two sensors are either aggregated without correction, leading to difficulty in noise refinement. Without GA, bi-directional information transfer is limited, failing to fully utilize the precise information from LiDAR. Combining both ideas allows effective information transfer without losing the advantages of corrected features from both sensors, reducing ambiguity, improving performance with lower std.

Table 2: Comparison with our KD variants. Best in **bold**, second best in underline.

	Methods	IoU=0.3		IoU=0.5	
		AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}
(a)	Ours-S (Multi-scale voxel KD loss)	73.1	82.4	38.7	64.9
(b)	Ours-S (L1 as voxel KD loss)	73.5	<u>82.8</u>	36.3	66.4
(c)	Ours-S (Struc. KD loss between L_l^T and L_l^S)	72.6	81.9	40.6	71.7
(d)	Ours-S (Struc. KD loss between R_l^T and L_l^S)	<u>74.8</u>	82.7	<u>41.2</u>	<u>71.9</u>
	Ours-S	**75.6**	**83.3**	**43.3**	**72.4**

Variant of KD Losses. We experiment with variations of the voxel KD loss as in (a) and (b) in Table 2. In (a), the performance degrades when applying the loss at a multi-scale level, indicating that distilling information from the low-level voxel features of the teacher, where radar information has not been fully incorporated, adversely affects performance. In (b), the performance also decreases when replacing the cosine similarity with the L1 loss, suggesting that introducing slight flexibility through cosine function is more beneficial than strictly aligning all feature values. Furthermore, we explore different teacher voxel features utilized in the structural KD loss in (c) and (d). Applying KL divergence solely to LiDAR teacher voxel resulted in a significant performance decrease while applying it only to radar teacher voxel showed a minor decrease. This highlights that distilling both inter-/intra-modal structural knowledge from 4D radar and LiDAR voxels is most beneficial for enhancing the LiDAR-only student model.

Objectness Score and Feature Visualization. We visualize the feature activation, objectness score map, and detection results under various weather conditions in Fig. 1. The feature activation map is calculated by densifying the sparse voxel features, finding maximum values along the channel dimension, and applying the average function along the Z-axis. Fig. 1 (b-g) are all visualized with the same numerical scale, without any post-processing. The LiDAR-based baseline model exhibits generally weak feature activation in regions where objects are present, as shown in (b). However, after computing the reliability score as in (b) and performing global aggregation with 4D radar followed by local aggregation, as shown in (f), significant activation is observed in regions where objects are present in the teacher feature, while suppressing the areas where activation was observed in LiDAR for the background. The visualization of the feature activate map reveals that our prompt learning method effectively integrates LiDAR and radar information, resulting in increased activation at locations where objects are present. Furthermore, in (g), it can be observed that the features from the student confidently activate at locations where objects are present, which validates the effectiveness of our four-level knowledge distillation method. This aspect is also observed in the objectness score maps derived from the classification maps. As depicted in (h-j), while the teacher model exhibits less confusion regarding where to focus compared to the baseline, the student model demonstrates a clear understanding of where to concentrate, accurately predicting the bounding box of the object.

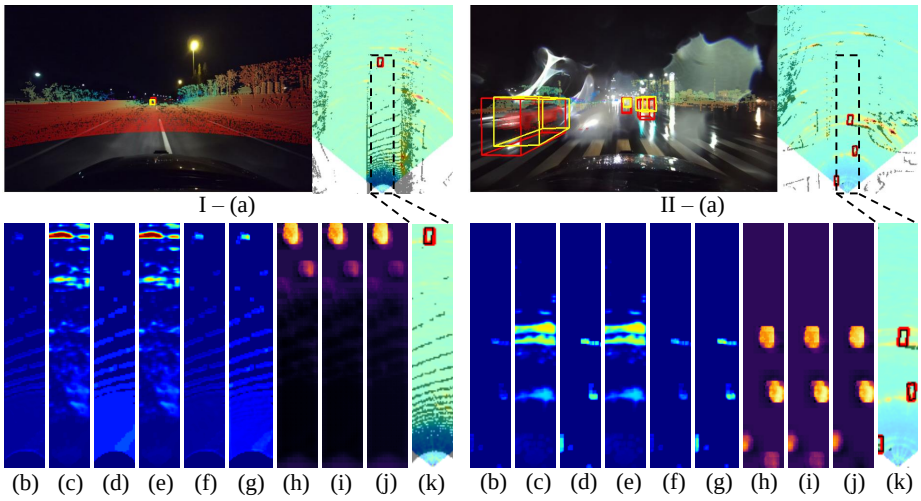

Fig. 1: (a) Visualized detection results of Ours-S. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. (b-g) Visualized feature activation of (b) L_1^T, (c) R_1^T, (d) $\phi(L_1^T)$, (e) $\widetilde{R}_1^T$, (f) $\hat{L}_1^T$ and (g) L_1^S. (h-j) Visualized objectness score map of (h) baseline (RTNH* [9]), (i) teacher (Ours-T) and (j) student (Ours-S). (I) Normal and (II) adverse (rain) conditions. Best viewed when zoomed in with colors.

Limitations. In adverse weather conditions, where LiDAR is noisy or sparser regarding objects, our LiDAR-based student model robustly detects objects. However, in scenarios when there is no input LiDAR data at all, our model naturally results in lower performance compared to model that utilize radar.

4 Additional Qualitative Results

We provide additional qualitative comparisons between our two models, the top three performing models following ours (BEVFusion* [7], RTNH* [9], VoxelNext [1]) and the 4D radar-only-based model (RTNH [9]). Results under various weather conditions, including normal, overcast, fog, rain, sleet, light snow, and heavy snow, are displayed in Fig. 2 to 8, respectively. We qualitatively demonstrate that our teacher model accurately detects objects in 3D across various weather conditions, including both normal and adverse conditions, surpassing competing models. Moreover, the student model, which solely relies on LiDAR, demonstrates robust performance across diverse weather conditions, even in scenarios with sparse LiDAR data, such as rain or sleet. Impressively, it outperforms the teacher model in all-weather conditions.

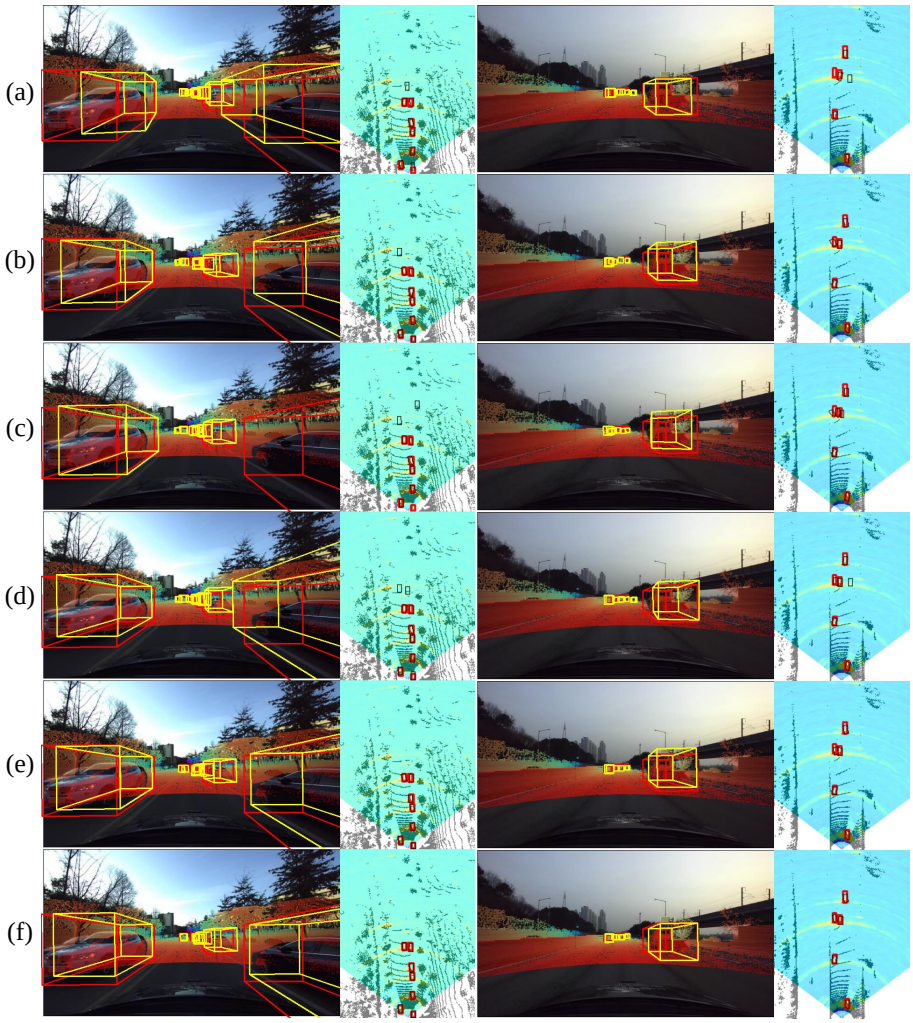

Fig. 2: Additional qualitative results under "normal" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

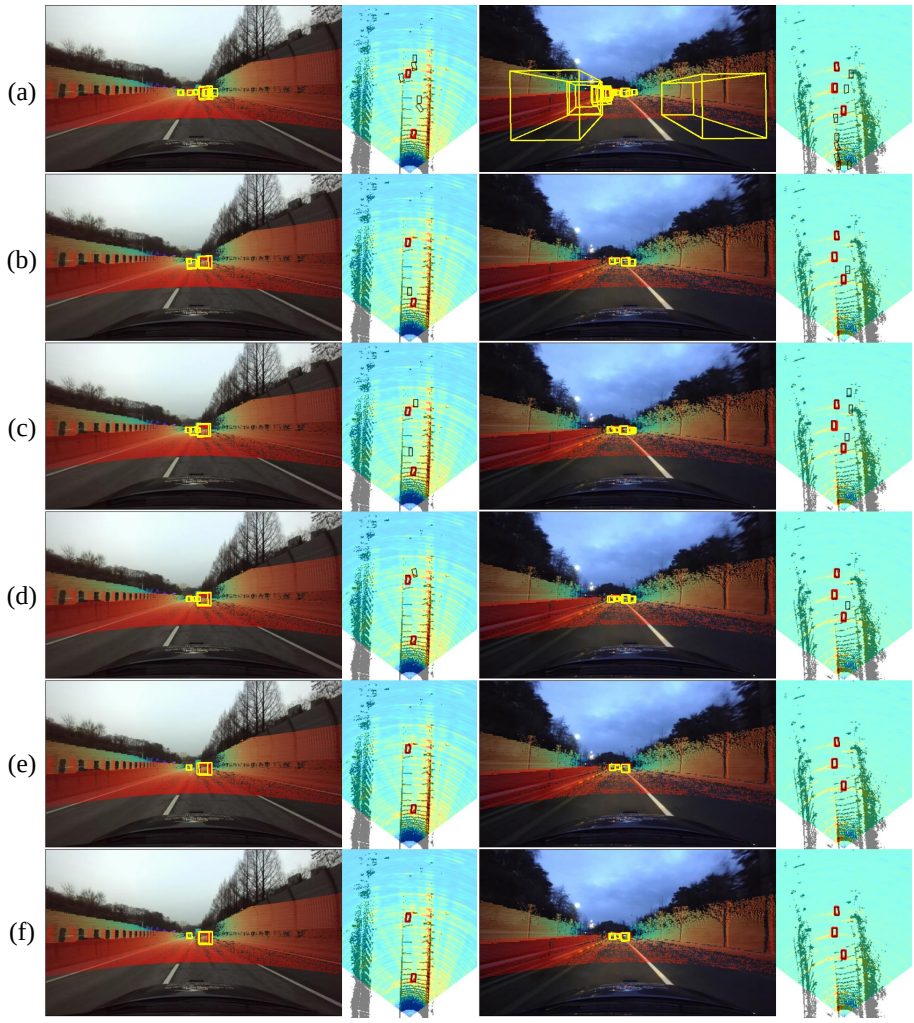

Fig. 3: Additional qualitative results under "overcast" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

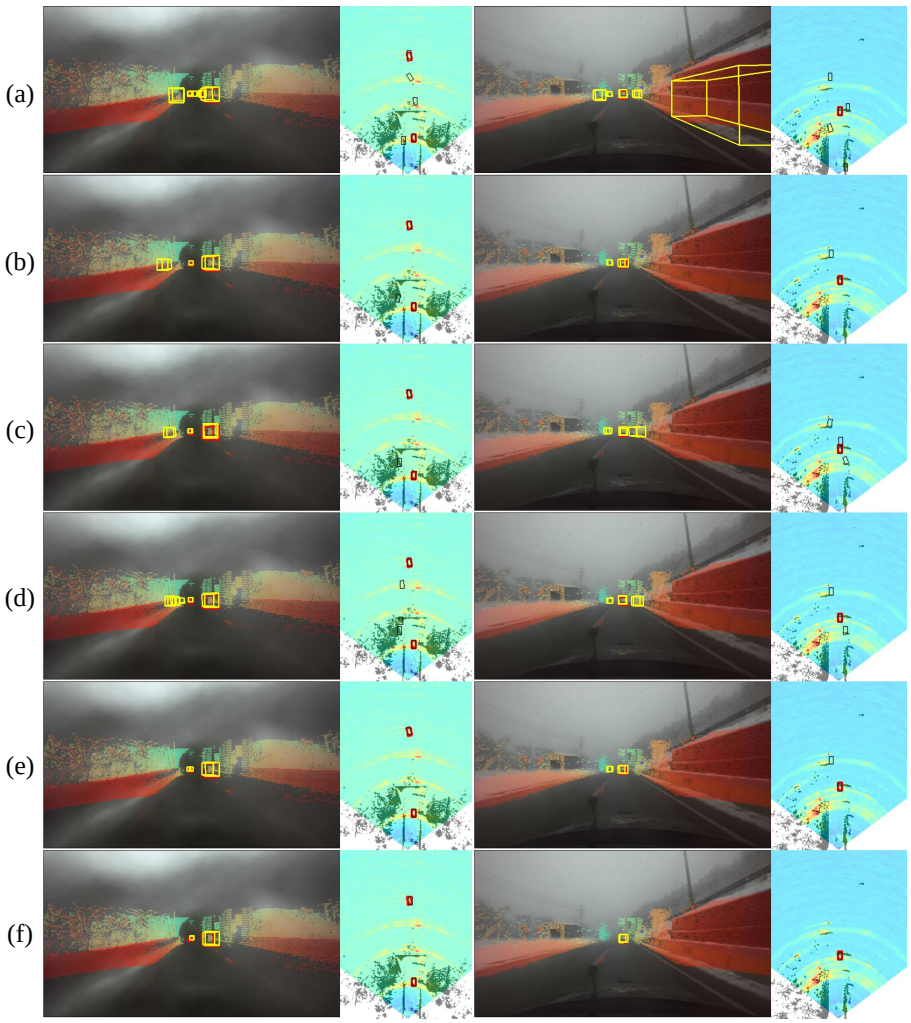

Fig. 4: Additional qualitative results under "fog" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

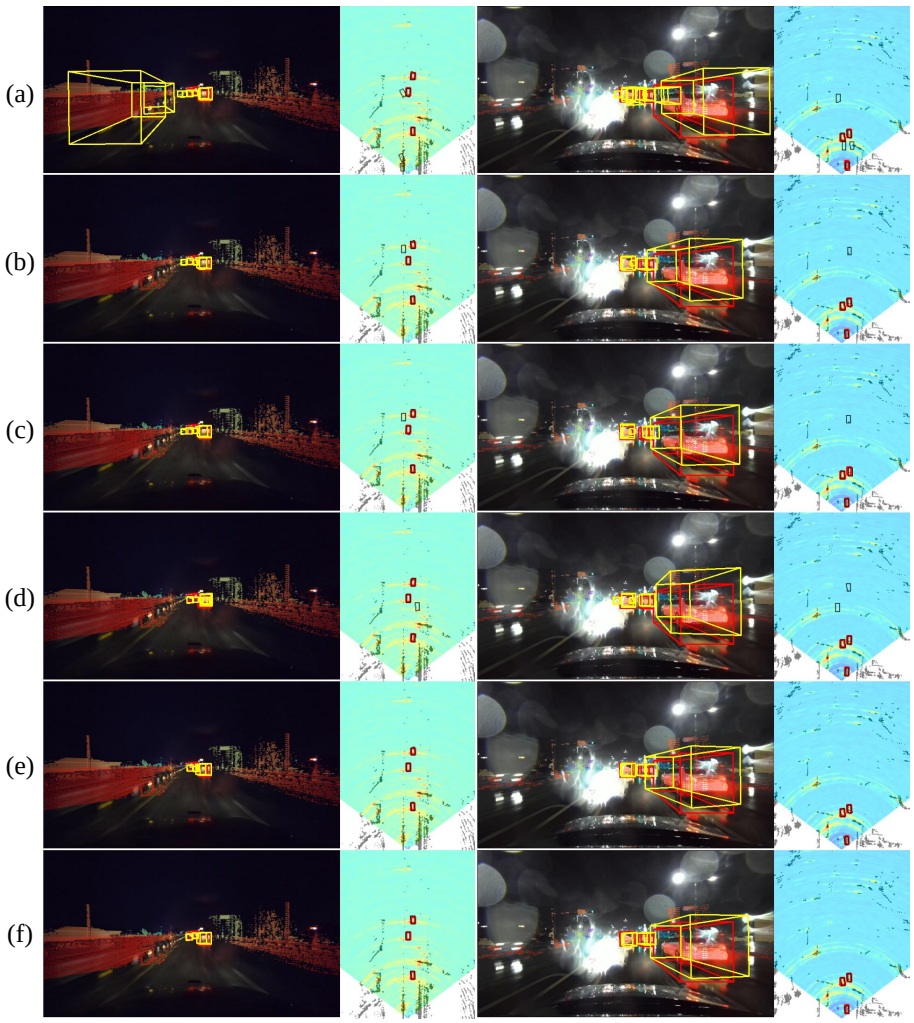

Fig. 5: Additional qualitative results under "rain" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

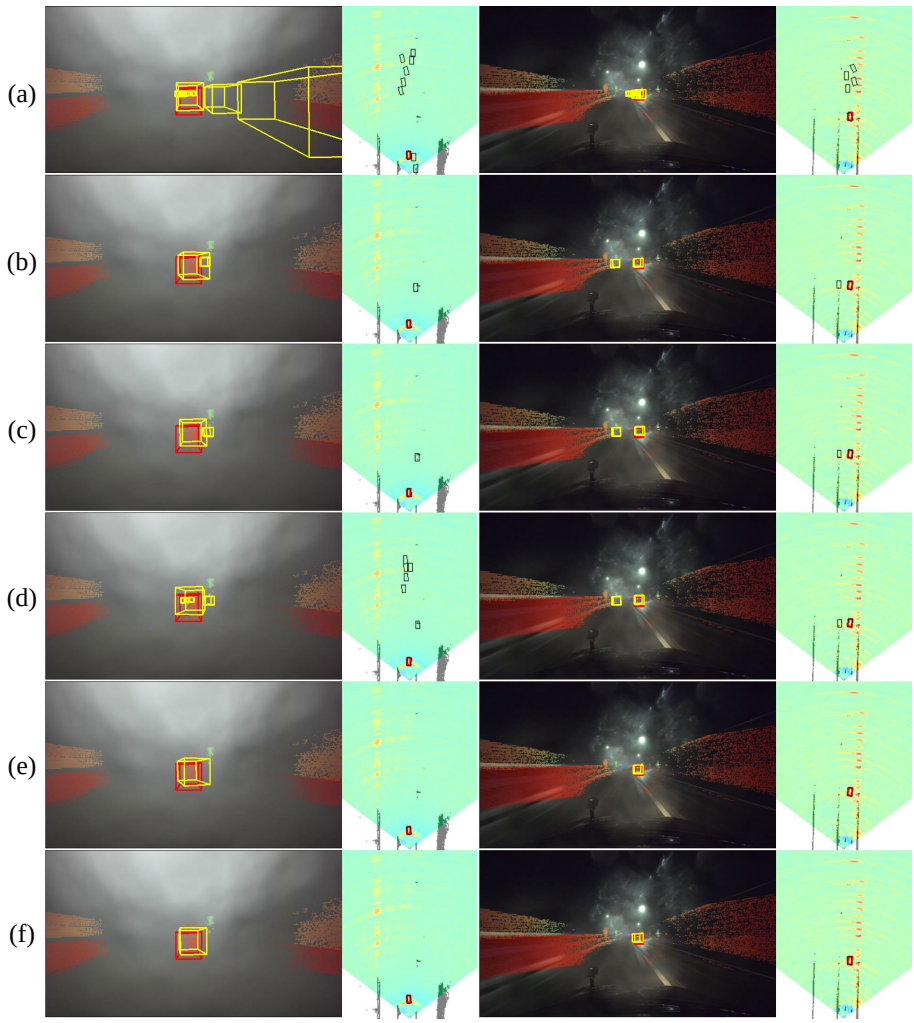

Fig. 6: Additional qualitative results under "sleet" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

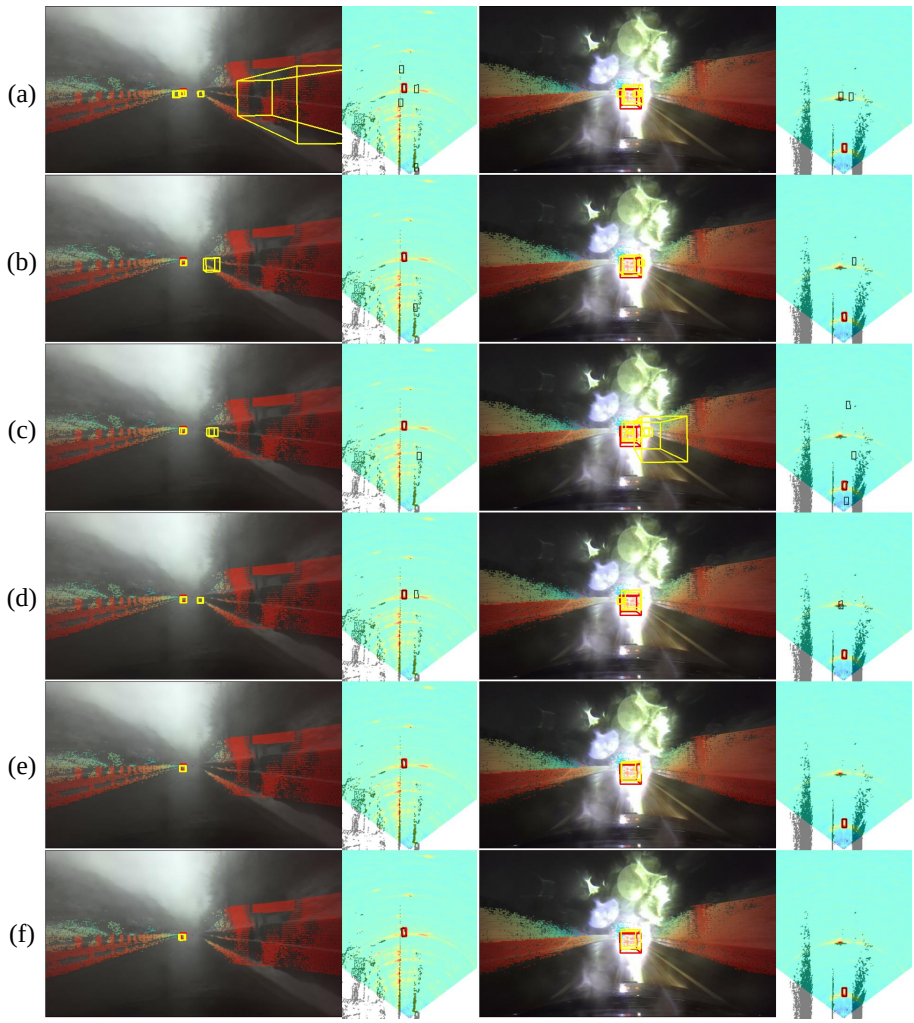

Fig. 7: Additional qualitative results under "light snow" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

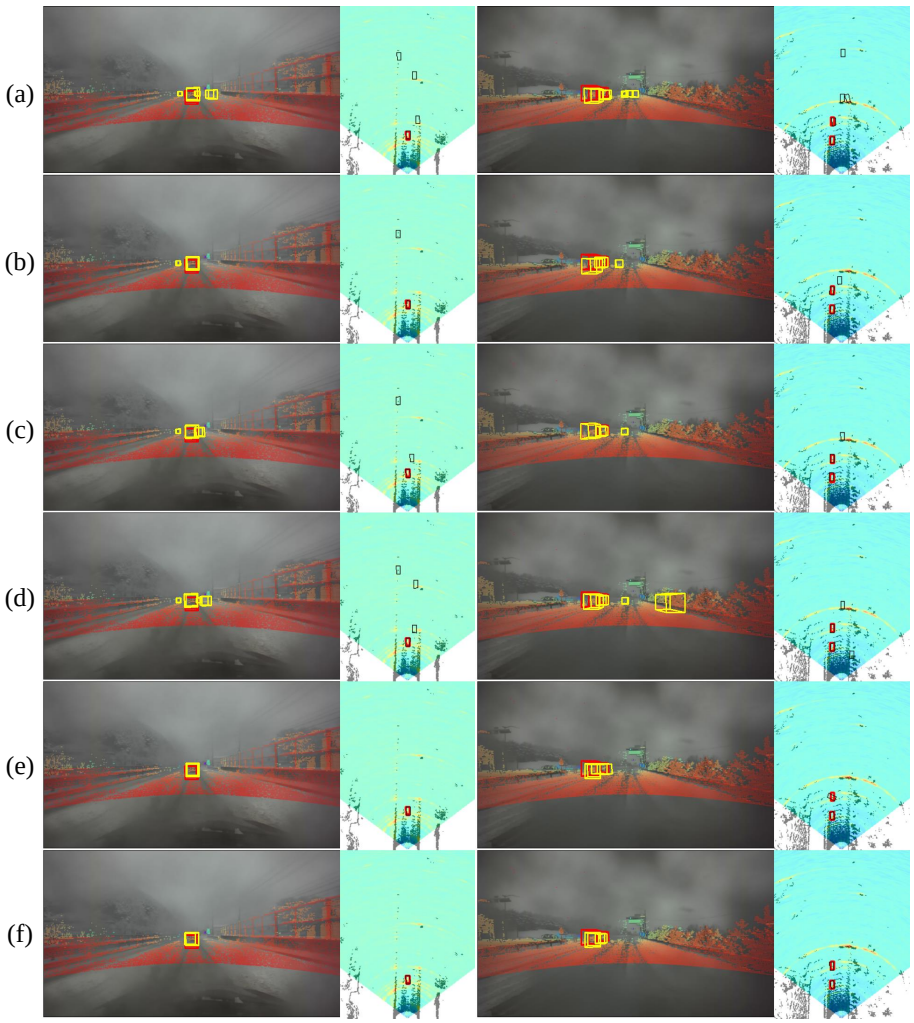

Fig. 8: Additional qualitative results under "heavy snow" conditions are presented. In the range view, images are accompanied by projected LiDAR data, with ground truth (GT) boxes marked in red and predicted boxes in yellow. In the bird's-eye-view, top-view LiDAR and 4D radar heatmaps are displayed, with GT boxes in red and predicted boxes in black. Each row corresponds to results from different methods: (a) RTNH [9], (b) RTNH* [9], (c) VoxelNext [1], (d) BEVFusion* [7], (e) Ours-T, and (f) Ours-S. Best viewed when zoomed in with colors.

References

1. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21674–21683 (June 2023)
2. Graham, B., van der Maaten, L.: Submanifold sparse convolutional networks. arXiv preprint arXiv:1706.01307 (2017)
3. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. p. 448–456. ICML'15, JMLR.org (2015)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
5. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3d: Towards robust and reliable 3d perception against corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19994–20006 (October 2023)
6. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12689–12697 (2018), https://api.semanticscholar.org/CorpusID:55701967
7. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE International Conference on Robotics and Automation (ICRA) (2023)
8. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=Skq89Scxx
9. Paek, D.H., Kong, S.H., Wijaya, K.T.: 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), https://openreview.net/forum?id=W_bsDmzwaZ7
10. Wang, L., Zhang, X., Xv, B., Zhang, J., Fu, R., Wang, X., Zhu, L., Ren, H., Lu, P., Li, J., Liu, H.: Interfusion: Interaction-based 4d radar and lidar fusion for 3d object detection. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 12247–12253 (2022). https://doi.org/10.1109/IROS47612.2022.9982123