CRN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception Supplementary Material

Youngseok Kim[1] Juyeb Shin[1] Sanmin Kim[1] In-Jae Lee[1] Jun Won Choi[2] Dongsuk Kum[1]
[1]KAIST [2]Hanyang University
{youngseok.kim, juyebshin, sanmin.kim, injaelee, dskum}@kaist.ac.kr, junwchoi@hanyang.ac.kr

A. Overview

This supplementary material provides additional details of architecture, qualitative and quantitative experimental results. We describe the notation of MDCA (Sec. B) and implementation details for experiments in the main paper (Sec. C). We further provide additional experimental results (Sec. D) and qualitative results (Sec. E).

B. Multi-modal Deformable Cross Attention

We adopt the deformable attention [29] and extend it for multi-modal feature maps, denoted as multi-modal deformable cross attention (MDCA).

Given an input queries z_q and flattened multi-modal BEV feature maps $\mathrm{x}_m = \{\mathbf{C}_I^{BEV}, \mathbf{C}_R^{BEV} \in \mathbb{R}^{C \times XY}\}$, let q index a query element and $p_q \in [0,1]^2$ be the normalized coordinates of the reference point for each query element q. The multi-modal deformable cross attention (MDCA) is defined as

$$\mathrm{MDCA}(\mathrm{z}_q, p_q, \mathrm{x}_m) = \sum_h^H \boldsymbol{W}_h \left[\sum_m^M \sum_k^K A_{hmqk} \cdot \boldsymbol{W}'_{hm} \mathrm{x}_m(\phi_m(p_q + \Delta p_{hmqk})) \right]. \tag{8}$$

h, m, k index the attention head H, multiple modalities $\{\mathbf{C}_I, \mathbf{C}_R\}$, and the number of sampling points K. $\boldsymbol{W}_h \in \mathbb{R}^{C \times C_v}$ is the output projection matrix at h^{th} head, and $\boldsymbol{W}'_{hm} \in \mathbb{R}^{C_v \times C}$ is the input value projection matrix at h^{th} head and modality m. We use $C_v = C/H$ following multi-head attention in Transformers [22]. Note that separated input value projection matrices $\boldsymbol{W}'_{hm}$ are used for each modality to make MDCA modality-specific and achieve robust fusion (*e.g.*, sensor failure case). Both A_{hmqk} and Δp_{hmqk} are obtained by linear projection over the input queries z_q, and the attention weight A_{hmqk} is normalized to modalities and sampling points as $\sum_m^M \sum_k^K A_{hmqk} = 1$. Function $\phi_m(p_q)$ scales the normalized coordinates p_q in case two modalities have different shapes.

The proposed multi-modal deformable attention module is designed to look over multi-modal feature maps and multiple sampling points. This can overcome spatial misalignment around reference points and enable adaptive fusion over modalities.

C. Implementation Details

This section provides the experimental settings for the main results and ablation studies.

C.1. Pre-processing and Hyper-parameters

For the camera stream, the image backbone yields 4 levels of feature maps of stride 4, 8, 16, and 32, and we employ SECONDFPN [23], which concatenates output feature maps at stride 16. `nn.Conv2d` and `nn.ConvTranspose2d` are used for downsampling and upsampling. Given FPN feature maps, the depth distribution network outputs D size depth bins. We use uniform discretization with a depth range of $[2.0, 58.0]m$ and bin size of $0.5m$, resulting in $D = 112$.

As stated in the main paper, we first project point cloud into an image coordinate system while preserving its depth and features for radar stream. Note that the projection matrix for radar point projection corresponds to the image stream. Next, we voxelize radar points in the frustum coordinate system (d, u, v) to have the same size with an image frustum feature. Taking into account the sparsity and accuracy of radar, we use $8\times$ downsampled pillar canvas and further extract pillar features using SECOND backbone, which yields 3 levels of feature maps of stride of 1, 2, and 4. Finally, SECONDFPN is employed to pillar feature maps to output $16\times$ downsampled size in the image width direction and to have $D = 112$ in a depth direction.

We use `multi_scale_deform_attn` implementation from MMCV [4] for deformable cross attention in Multi-modal Feature Aggregation (MFA). Specifically, we use 6 layers of MFA, 8 attention heads, and 4 sampling points for MFA in our experiments. MFA is applied to the single-frame camera and radar inputs and produces a fused

configs	ResNet-50/101	ConvNeXt-B
optimizer	AdamW	AdamW
base learning rate	2e-4	1e-4
backbone learning rate	2e-4/1e-4	5e-5
weight decay	1e-4	1e-2
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$
batch size	64/32	16
training epochs	24	24
lr schedule	step decay	step decay
gradient clip	5	5
stochastic depth [8]	None	0.4
layer scale [21]	None	1.0

Table 1. Training settings for the main results.

BEV feature map. After, fused BEV feature maps from the previous T timestamps are aligned to the current timestamp and concatenated. We use $T = 3$ for the submission and $T = 1$ for ablation studies and note that future frames are not used.

Following standard practices in monocular 3D object detection [9, 13], we set perception range $[-51.2, 51.2]m$ with a pillar size of $(0.2, 0.2)m$ and a downsampling factor of 4. As a result, the size of the BEV feature map is 128×128.

For the BEV segmentation task, the perception range is set to $[-50.0, 50.0]m$ in both $X-$ and $Y-$axis centered around the ego vehicle, following previous works [20, 7, 26, 2]. The resolution of the final output is $0.5m$, resulting in 200×200 grid map.

C.2. Training Settings

All models are trained for 24 epochs with AdamW [17] optimizer in an end-to-end manner. Image backbones are pre-trained on ImageNet [5]. In Table 1, we provide ResNet [6] and ConvNeXt [15] training settings used for our main results.

For image and radar data augmentation (in perspective view), we use resize, crop, and horizontal flipping augmentation following standard practices [9, 13]. We discard rotation augmentation since the rotation can have an adverse effect when collapsing the height dimension in radar-assisted view transformation (RVT). Note that the same data augmentation is applied to the image and radar in the perspective view.

For BEV augmentation, we use random flipping along X and Y axis, global rotation between $[-\pi/8, \pi/8]$, and global scale between $[0.95, 1.05]$. BEV data augmentation is applied to the BEV feature map and ground truth boxes correspondingly. Note that ground-truth sampling augmentation (GT-AUG) [23] is not used in our experiments, and we leave GT-AUG for a multi-modal setting [3, 25] as future work.

C.3. Baselines for Ablation Studies

We conduct three baselines BEVDepth [13], CenterPoint [24], and BEVFusion [16] for camera-only, point-only, and camera-point fusion detectors. Note that CenterPoint and BEVFusion originally take LiDAR points as input and we replace LiDAR points $(x, y, z, intensity)$ to radar points $(x, y, z, RCS, Doppler)$ without network modification.

For BEVDepth, we use the official code[1] without class-balanced grouping and sampling (CBGS) [28] and exponential moving average (EMA).

For CenterPoint, we use MMDetection[2] implementation using PointPillar [11] backbone with $(0.2, 0.2, 8)m$ pillar size. Different from the official implementation, CBGS [28] and GT-AUG [23] are discarded for fair comparisons.

For BEVFusion, we use BEVDepth for obtaining the camera BEV feature map and CenterPoint-Pillar for point BEV feature maps and fuse them by a single 3×3 convolution layer following official implementation. Note that our BEVFusion may yield better performance since our implementation uses BEVDepth for the camera stream, while the original BEVFusion uses LSS [20].

C.4. Details of Long-Range Model

To analyze the performance of CRN over long perception ranges, we increase the perception range of baselines to $[-102.4, 102.4]m$. For camera streams, we increase the range of depth distribution from $[2.0, 58.0]m$ to $[2.0, 116.0]m$, and the number of depth bins is doubled to $D = 224$. For point streams, the range of point cloud and pillars are increased to correspond to the perception range. Note that we use the same pillar size $(0.2, 0.2)m$ and downsampling factor of 4, resulting in a 256×256 BEV feature map for all baselines.

For training and evaluating long-range models, we increase the 'class range' in nuScenes [1] twice to filter the ground truth and predictions. Particularly, the class range of car, truck, bus, trailer, and construction vehicle are $100m$, pedestrian, motorcycle, and bicycle are $80m$, traffic cone and barrier are $60m$. Moreover, nuScenes filters annotation that does not contain at least single LiDAR or radar point inside the 3D bounding box (denote as 'points in box filtering') for training and evaluation, but we disable this filtering for thorough analysis. Thus, some moving objects are visible on the image but do not have annotations (due to not enough points to label), and some static objects can have annotations but are not visible on the image (labeled on the previous timestamp but occluded on the current timestamp) in our setting. Although disabling point filtering may cause inconsistency between input data and annotation and

[1] https://github.com/Megvii-BaseDetection/BEVDepth
[2] https://github.com/open-mmlab/mmdetection3d

Method	Input	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	Bicycle	T.C.	Barrier	mAP
CenterPoint-P [24]	L	83.9	49.5	61.9	34.1	12.3	76.9	44.1	18.0	54.0	59.1	49.4
CenterNet [27]	C	48.4	23.1	34.0	13.1	3.5	37.7	24.9	23.4	55.0	45.6	30.6
CenterFusion [18]	C+R	52.4(+4.0)	26.5(+3.4)	36.2(+2.2)	15.4(+2.3)	5.5(+2.0)	38.9(+1.2)	30.5(+5.6)	22.9(-0.5)	56.3(+1.3)	47.0(+1.4)	33.2(+2.6)
CRAFT-I [10]	C	52.4	25.7	30.0	15.8	5.4	39.3	28.6	29.8	57.5	47.8	33.2
CRAFT [10]	C+R	69.6(**+17.2**)	37.6(**+11.9**)	47.3(**+17.3**)	20.1(+4.3)	10.7(+5.3)	46.2(+6.9)	39.5(**+10.9**)	31.0(+1.2)	57.1(-0.4)	51.1(+3.3)	41.1(+7.9)
BEVDepth [13]	C	55.3	25.2	37.8	16.3	7.6	36.1	31.9	28.6	53.6	55.9	34.8
CRN	C+R	73.6(**+18.3**)	44.5(**+19.3**)	55.6(**+17.8**)	22.0(+5.7)	15.4(+7.8)	50.2(**+14.1**)	54.7(**+22.8**)	48.9(**+20.3**)	61.4(+7.8)	63.8(+7.9)	49.0(**+14.2**)

Table 2. Per-class comparisons on nuScenes `val` set. 'C.V.', 'Ped.', 'M.C.', and 'T.C.' denote construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. CenterNet [27], CRAFT-I [10], and BEVDepth [13] are camera baselines of CenterFusion [18], CRAFT [10], and CRN. CenterPoint-P and BEVDepth results are from MMDetection3D and the official code.

# Frames	NDS	mAP	mATE	mAOE	mAVE
1	50.3	42.9	0.519	0.577	0.520
2	54.5	46.0	0.495	0.538	0.350
3	55.7	47.3	0.480	0.507	0.342
4	56.0	48.1	0.474	0.541	0.328
5	56.4	48.4	0.469	0.515	0.345

Table 3. Ablation of temporal frames.

# Top-K	AP	ATE	AOE	AVE	FPS
1024	49.8	0.399	0.216	0.371	14.1
2048	52.4	0.382	0.202	0.352	14.0
4096	54.0	0.367	0.194	0.340	14.0
8192	54.6	0.362	0.191	0.352	13.8
All	56.9	0.325	0.158	0.298	11.5

Table 4. Ablation of sparse aggregation.

harms performance during training, all methods are trained and evaluated using the same setting for a fair comparison. We find that the inference speed of CenterPoint [24] with radar input is much faster than LiDAR input, assuming that the sparsity of radar points can highly benefit from voxelization and sparse convolution [23].

D. Additional Experimental Results

D.1. Per-Class Analysis

In Table 2, we compare the performance improvement of camera-radar methods over camera-only baselines. For fair comparisons, we report 256×704 and R50 models for BEVDepth and CRN. Corresponds to results on CRAFT [10], metallic and frequently appeared on road classes (car, truck, bus, and motorcycle) gain significant improvements. Different from CRAFT, ours also shows a huge improvement in non-metallic classes (pedestrian, bicycle, traffic cone, and barrier). Moreover, we find that the performance gain of using radar is much more significant on ours than other fusion methods. Considering the performance difference of camera baselines are not significant, results in Table 2 demonstrate that the design of fusion methods greatly affects the performance.

D.2. Design Decisions

We study architecture designs that affect the performance of CRN to provide insights on the proposed sensor fusion framework.

Temporal Frames. We accumulate multiple BEV feature maps on channel dimension by concatenation and aggregate them by a few convolutional layers before feeding them to the BEV backbone. We find that the time interval of 1 second yields a better performance than 0.5 second proposed in previous approaches [14, 13]. Compared to temporal stereo methods [12, 19], ours does not require sequential data input for obtaining the BEV feature map; thus, using an arbitrary number of BEV feature maps does not increase latency. We note that BEV feature maps on previous timestamps are obtained without gradients during training following standard practices.

As shown in Table 3, using multiple temporal frames significantly improves mAP, mATE, and mAVE. Corresponding to results on recent approaches using temporal BEV feature maps [19], a larger number of frames consistently yields better performance. However, we observe the unstable orientation error (mAOE), suggesting room for improvement in utilizing BEV feature maps, and we leave this as future work. As the performance gain is saturated on four frames, we decide to use four frames considering computation time and memory during training.

Sparse Aggregation. In Table 4, we ablate the number of N_k feature grids on sparse aggregation settings. Note that the total number of BEV feature grids is $N = 256 \times 256 = 65536$ in our long-range setting and we report the performance on *Car* class at $100m$ perception range. Since the computational complexity of sparse aggregation $\mathcal{O}(2N_k + N_k K)$ is linear to sparse input queries N_k, using a small set of features for MFA significantly reduces the computation of Multi-modal Deformable Cross Attention (MDCA). More specifically, using 4096 size queries reduce the latency of MFA by 76.4% ($21.01ms$ to $4.96ms$) on 256×256 size BEV grid. However, as the BEV feature map becomes sparse and discretized after top-k sampling, the performance is degraded. We find that the performance drops on True Positive metrics (*e.g.*, ATE, AOE, AVE) are more significant than AP, assuming that the classification network can maintain its performance, but the re-

gression network suffers from sparsely spread BEV features to regress objects' attributes.

E. Additional Qualitative Results

We show additional qualitative results of 3D object detection (long-range 256×704 and R50 model) and BEV segmentation (256×704 and R50 model).

To visualize 3D detection results for the range of $200m \times 200m$, we disable points in box filtering as described in Appendix C.4. As can be seen in Fig. 1, CRN is capable of detecting objects even at a very far distance under various and complex driving scenarios. Thanks to radar fusion, objects strongly occluded by other objects or hardly visible by low lighting are successfully detected by ours. Moreover, even if some objects do not have radar point returns, CRN can still detect them by image only. Failure cases of CRN are likely caused when objects are rare classes and do not without radar points (*e.g.*, construction vehicles behind wire mesh or trailers heavily occluded).

We further visualize BEV segmentation results in the range of $100m \times 100m$, following the same setting of previous works. As shown in Fig. 2, CRN is also capable of accurately predicting segmentation occupancy of drivable region and vehicle. Thanks to our camera-radar fusion framework to generate a semantically rich and spatially accurate feature map, our results show stable performance under various lighting and weather conditions. CRN can further predict occupancy of the vehicles with a complete shape both at nearby and faraway distances, even when the vehicles are partially visible. In terms of the drivable region, CRN can successfully predict the complex shapes of the road even under occlusions.

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 (CVPR)*, pages 11621–11631, 2020. 2

[2] Shaoyu Chen, Tianheng Cheng, Xinggang Wang, Wenming Meng, Qian Zhang, and Wenyu Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*, 2022. 2

[3] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 628–644, 2022. 2

[4] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[7] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15273–15282, 2021. 2

[8] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–661, 2016. 2

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

[10] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

[11] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 2

[12] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

[13] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 3

[14] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18, 2022. 3

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 2

[16] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[18] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1527–1536, 2021. 3

[19] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[20] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 194–210, 2020. 2

[21] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, 2021. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017. 1

[23] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337–3352, 2018. 1, 2, 3

[24] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021. 2, 3

[25] Wenwei Zhang, Zhe Wang, and Chen Change Loy. Exploring data augmentation for multi-modality 3d object detection. *arXiv preprint arXiv:2012.12741*, 2020. 2

[26] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13760–13769, 2022. 2

[27] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 3

[28] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. In *arXiv preprint arXiv:1908.09492*, 2019. 2

[29] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

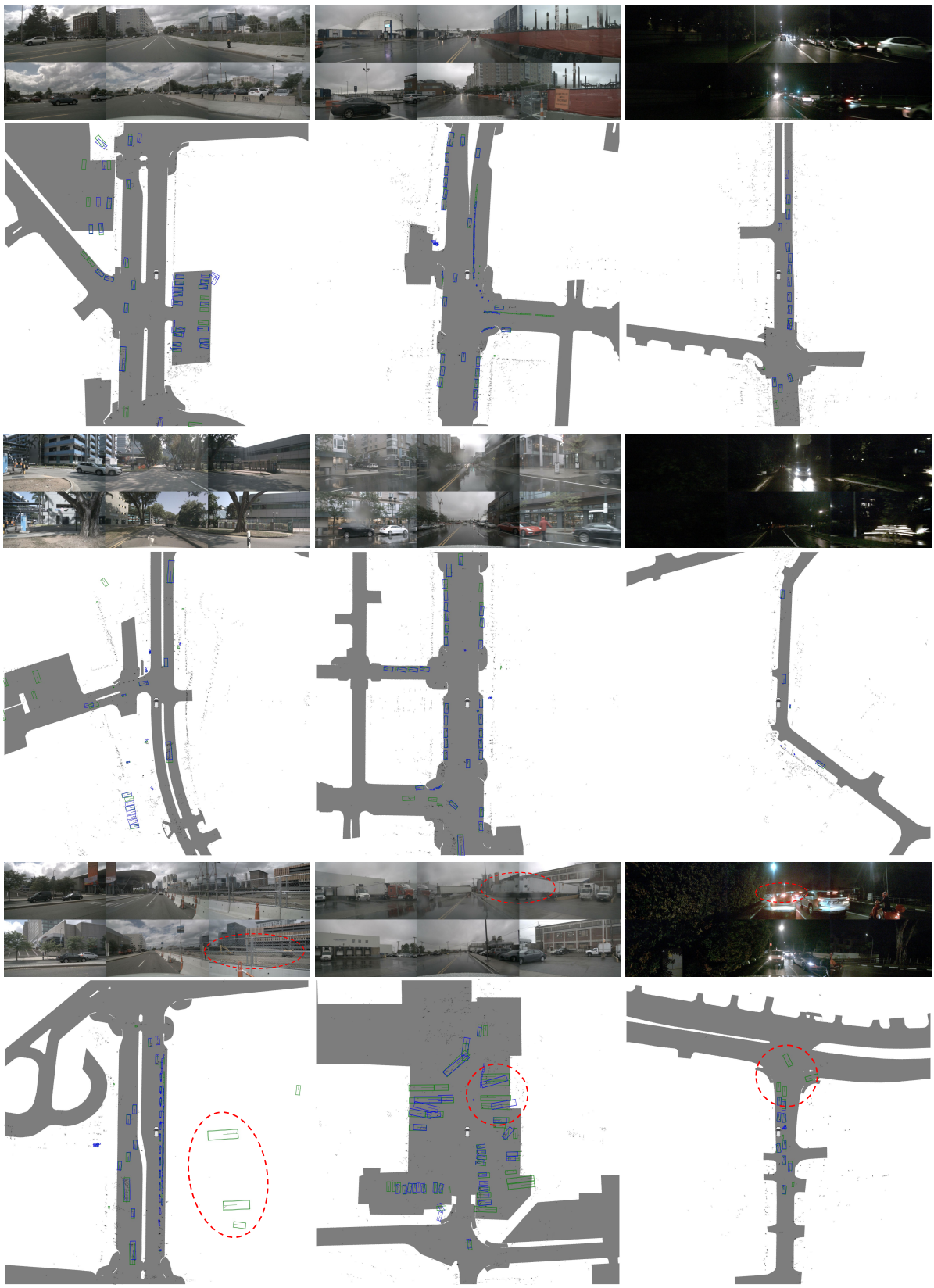

Figure 1. Additional qualitative results of 3D object detection on nuScenes `val` set: from left to right, day, rainy, and night scenarios. Green boxes are ground truths, blue boxes are our prediction results, and black dots are radar points. We also show the failure cases and highlight them with red circles on the bottom row. Ground truth maps on the background are used for visualization. Best viewed in color with zoom in.

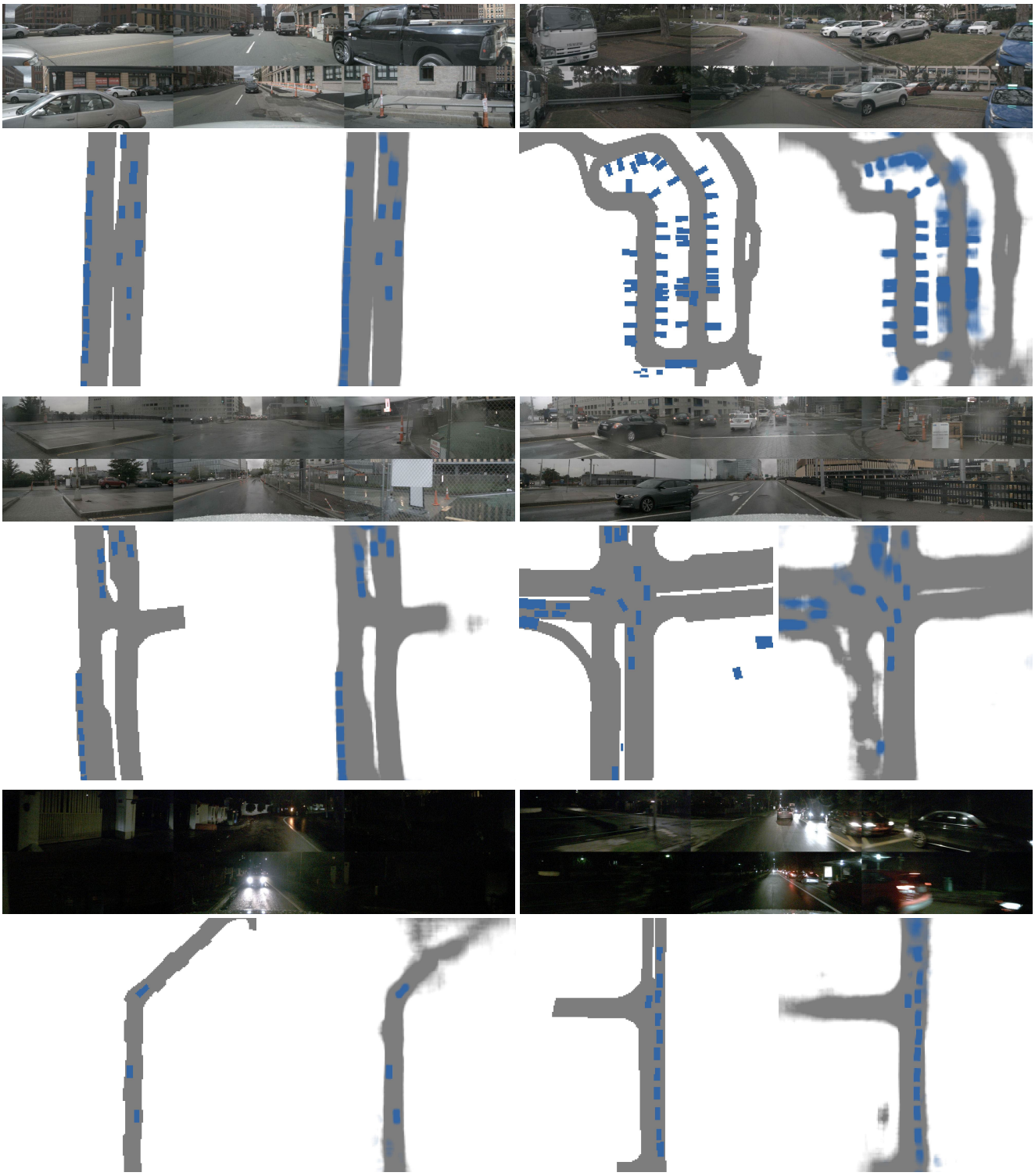

Figure 2. Qualitative results of BEV segmentation on nuScenes `val` set on various road shapes and weather conditions: from top to bottom, day, rainy, and night scenarios. Images on the top are the six camera views surrounding the vehicle, the bottom left is ground truth, and the bottom right is our prediction results. Best viewed in color with zoom in.