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

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

Abstract

Autonomous driving requires a robust and reliable 3D perception system that includes 3D object detection, tracking, and segmentation. Although recent low-cost camera-based approaches have shown promising results, they are susceptible to poor illumination or bad weather conditions and have a large localization error. Hence, fusing camera with low-cost radar, which provides precise long-range measurement and operates reliably in all environments, is promising but has not yet been thoroughly investigated. In this paper, we propose Camera Radar Net (CRN), a novel camera-radar fusion framework that generates a semantically rich and spatially accurate bird's-eye-view (BEV) feature map for various tasks. To overcome the lack of spatial information in an image, we transform perspective view image features to BEV with the help of sparse but accurate radar points. We further aggregate camera and radar feature maps in BEV using multi-modal deformable attention designed for adaptive fusion given spatially misaligned and ambiguous multi-modal inputs. CRN with a real-time setting operates at 20 FPS while achieving comparable performance to LiDAR detectors on nuScenes, and even outperforms at a $100m$ perception range. Moreover, CRN with offline setting yields 58.3% NDS, 51.5% mAP at 7 FPS and is ranked first among all camera and camera-radar 3D object detectors. The code will be made publicly available soon.

1 Introduction

Accurate and robust 3D perception system is crucial for many applications such as autonomous driving and mobile robot. For efficient 3D perception, obtaining a reliable bird's eye view (BEV) feature map from sensor inputs is necessary since various downstream tasks can be operated on BEV space (*e.g.*, object detection & tracking (Yin et al., 2021), map segmentation (Zhou & Krähenbühl, 2022), trajectory prediction (Hu et al., 2021), and motion planning (Philion & Fidler, 2020)). Another important ingredient for deploying 3D perception to the real world is to build a system that relies less on high-cost, high-maintenance, and low-reliable LiDAR sensors. Apart from the drawbacks of LiDAR, 3D perception system is required to identify semantic information on the road (*e.g.*, traffic lights, road sign) that can be easily leveraged by camera. In addition to the need for rich semantic information, detecting distant objects is essential, and this can be advantaged from radar.

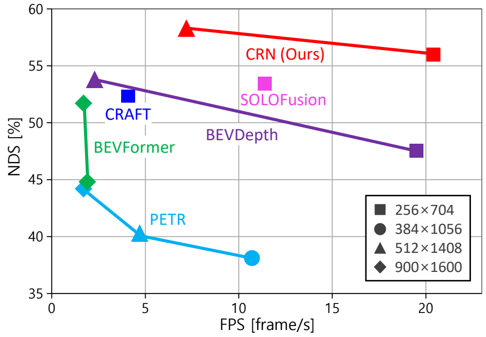

Figure 1: FPS vs. accuracy on nuScenes `val` set. We show that fusing radar can significantly boost camera-only method with marginal computational cost. CRN outperforms all methods with much faster speed. See Table 1 and Fig. 5 for more details.

Recently, camera-based 3D perception in BEV (Philion & Fidler, 2020; Reading et al., 2021; Huang et al., 2021) has drawn great attention. Thanks to rich semantic information in dense image pixels, camera approaches can distinguish objects even at a far distance. Despite the advantage of cameras in cost, localizing accurate position of objects from a monocular image is naturally a challenging ill-posed problem. Moreover, cameras can be significantly affected by illumination conditions (*e.g.*, glare, low-contrast, or low-lighting) due to the nature of the passive sensor. To address this, *we aim to generate a BEV feature map using a camera with the help of a cost-effective range sensor, radar*.

Radar has advantages not only in cost but high-reliability, long-range perception (up to $200m$ for typical automotive radar (ars)), robustness in various conditions (*e.g.*, snow, fog, or rain), and providing velocity estimation from a single measurement. However, radar also brings its challenges such as sparsity (typically $180\times$ fewer than LiDAR points per single frame in nuScenes (Caesar et al., 2020)), noisy and ambiguous measurements (false negatives by low resolution, accuracy, or low radar cross-section and false positives by multi-path or clutters). As a result, previous camera-radar fusion methods using late fusion strategies that fuse detection-level results (Göhring et al., 2011; Cho et al., 2014) fail to fully exploit the complementary information of sensors, thus having limited performance and operating environment. Despite the huge potential of learning-based fusion, only a few studies (Kim et al., 2023; Nabati & Qi, 2021; Kim et al., 2020b) explore camera-radar fusion in autonomous driving scenarios.

To put the aforementioned advantages and disadvantages of camera and radar in perspective, camera-radar fusion should be capable of following properties to fully exploit the complementary characteristics of each sensor. *First*, camera features should be accurately transformed into BEV space in terms of spatial position. *Second*, the fusion method should be able to handle the spatial misalignment between feature maps when aggregating two modalities. *Last but not least*, all mentioned above should be adaptive in order to tackle noisy and ambiguous radar measurements.

To this end, we design a novel two-stage fusion method for BEV feature encoding, *Camera Radar Net (CRN)*. The key idea of the proposed method is to generate *semantically rich and spatially accurate BEV feature map* by fusing complementary characteristics of camera and radar sensors. In particular, we first transform camera image features in perspective view into BEV, not solely relying on estimated depth information but using radar, named *radar-assisted view transformation (RVT)*. Since transformed image features in BEV is not completely accurate, following *multi-modal feature aggregation (MFA)* consecutively encodes the multi-modal feature maps into a unified feature map using an attention mechanism. We conduct extensive experiments on nuScenes and demonstrate that our proposed method can generate a fine-grained BEV feature map to set the new state-of-the-art while maintaining high efficiency, as shown in Fig 1.

The main contribution of our works are three-fold:

- **Accurate.** CRN achieves LiDAR-level performance on 3D object detection task only using cost-effective camera and radar.
- **Robust.** CRN maintains robust performance even if one of the single sensor inputs is entirely off, which allows the fault-tolerant system.
- **Efficient.** CRN requires marginal extra cost for significant performance improvement, which enables long-range perception in real-time.

2 Related Work

Camera-based 3D Perception. Thanks to well-established 2D object detection methods (Ren et al., 2015; Zhou et al., 2019; Tian et al., 2019) on perspective view images, early approaches extend 2D detector to 3D detector by additionally estimating object's depth (Simonelli et al., 2019; Wang et al., 2021b;a; 2022), then transforming object center. DD3D (Park et al., 2021) improves detection performance by pre-training depth estimation task on depth dataset (Guizilini et al., 2020). Although a simple and intuitive approach, the view discrepancy between input feature space (perspective view, PV) and output space (bird's-eye-view, BEV) restricts the network from extending to other tasks.

Recent advances in camera-based perception exploit view transformation. Geometry-based methods (Philion & Fidler, 2020; Reading et al., 2021; Li et al., 2023b; Park et al., 2022) explicitly estimate the depth distribution of each image feature and transform it by outer product. BEVDepth (Li et al., 2023b) empirically shows that training depth distribution with auxiliary pixel-wise depth supervision improves the performance, which corresponds to the results of DD3D (Park et al., 2021). Learning-based methods (Li et al., 2022c; Zhou & Krähenbühl, 2022; Jiang et al., 2023;

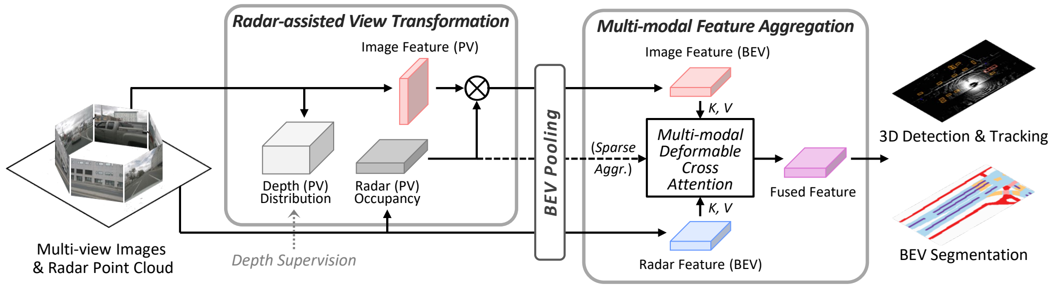

Figure 2: The overall architecture of the proposed Camera Radar Net. Given multi-view images and radar points, modality-specific backbones extract features in each view. First, image context features in perspective view are transformed into a bird's-eye-view with the help of radar measurements by Radar-assisted View Transformation (RVT). After, Multi-modal Feature Aggregation (MFA) adaptively aggregates image and radar feature maps to generate semantically rich and spatially accurate bird's-eye-view representation.

Lu et al., 2022) implicitly model the mapping function from PV to BEV using multi-layer perceptron (MLP) (Roddick & Cipolla, 2020; Saha et al., 2022) or cross-attention (Li et al., 2022c).

Obtaining a BEV feature map allows the framework to be easily extended to various downstream tasks performed on BEV space, such as 3D detection and tracking (Li et al., 2023b), segmentation (Zhou & Krähenbühl, 2022), and prediction (Philion & Fidler, 2020). However, camera-only methods have low localization accuracy due to the absence of distance information in image and are sensitive to lighting or weather conditions. Moreover, achieving LiDAR-level performance using camera-only methods requires large image input and backbone, which is slow and not applicable for real-time application.

Camera-Point 3D Perception. Fusing complementary information of camera image and range measurement is a promising and active research topic, and the view discrepancy between sensors is regarded as a bottleneck for multi-modal fusion. A line of approaches handles discrepancy by projecting 3D information to a 2D image (*e.g.*, points (Vora et al., 2020; Li et al., 2022b), proposals (Kim et al., 2020b; Bai et al., 2022), or prediction results (Pang et al., 2020)) and gathering information around the projected region. Some camera-radar fusion methods (Lin et al., 2020; Long et al., 2021) attempt to improve depth estimation by projecting radar points to the image.

On the other hand, another line of work lift 2D image information into 3D. Early studies in 3D detection (Qi et al., 2018; Kim et al., 2023) detect 2D or 2.5D object proposals and then lift them into 3D space to fuse with point data; however, this object-level fusion is difficult to be generalized to other tasks in BEV. Thanks to advances in monocular BEV approaches, recent fusion approaches extract image and point feature maps in unified BEV space and then fuse feature maps by element-wise summation or concatenation, assuming multi-modal feature maps are spatially well aligned. After, the fused BEV feature map is used in various perception tasks such as 3D detection (Yoo et al., 2020; Liang et al., 2022; Li et al., 2022a), BEV segmentation (Harley et al., 2022), or multi-task (Liu et al., 2022b; Zhou et al., 2023). However, although unique characteristics of a camera (*e.g.*, inaccurate BEV transformation) and radar (*e.g.*, sparsity and ambiguity), previous camera-radar fusion less considers them. Our proposed CRN focuses on fusing multi-modal feature maps considering the characteristics of each sensor thoroughly to have the best of both worlds.

3 Camera Radar Net

In this paper, we propose a camera radar fusion framework to produce a unified BEV representation given multi-view images and radar points, as illustrated in Fig 2. In Sec. 3.2, we introduce a method to transform image features with radar, then a multi-modal feature aggregation method in Sec. 3.3. Finally, generated BEV feature map is used for downstream tasks in Sec. 3.4.

3.1 Preliminary

Monocular 3D Approaches. The crux of monocular 3D perception is *how to construct accurate 3D (or BEV) information from 2D features*, which can be categorized into two groups. Geometry-based approaches (Philion & Fidler, 2020; Reading et al., 2021; Huang et al., 2021; Li et al., 2023b) predict depth $\mathbf{D}$ as an explicit intermediate representation and transform features $\mathbf{F}$ in perspective view (u, v) into frustum view (d, u, v) then 3D (x, y, z) by:

$$\mathbf{F}_{3D}(x, y, z) = \mathcal{M}(\mathbf{F}_{2D}(u, v) \otimes \mathbf{D}(u, v)), \tag{1}$$

where $\mathcal{M}$ denotes view transformation module (*e.g.*, Voxel Pooling (Liu et al., 2022b; Li et al., 2023b)) and $\otimes$ denotes outer product. Meanwhile, learning-based approaches (Li et al., 2022c; Liu et al., 2022a; Wang et al., 2022; Zhou & Krähenbühl, 2022) implicitly model 3D to 2D projection utilizing mapping networks as:

$$\mathbf{F}_{3D}(x, y, z) = f(P_{xyz}, \mathbf{F}_{2D}(u, v)), \tag{2}$$

where f denotes mapping function between perspective view and BEV (*e.g.*, multi-layer perceptron (MLP) (Saha et al., 2022) or cross-attention (Li et al., 2022c)), and P_{xyz} is voxels in 3D space. Although the approaches are different, the key is to obtain spatially accurate 3D features $\mathbf{F}_{3D}(x, y, z)$ through implicit or explicit transformation. In this paper, we aim to improve transformation using radar measurement explicitly.

Radar Characteristics. Radar data can have various representations (*e.g.*, 2-D FFT (Lin et al., 2018), 3D Tensor (Major et al., 2019; Kim et al., 2020a), point cloud (Caesar et al., 2020; Meyer & Kuschk, 2019)). Radar point cloud has a similar representation to LiDAR, but their sensor characteristics are different in terms of resolution and accuracy (ars). Moreover, due to the nature of the operating mechanism of radar (Johnson & Dudgeon, 1992; Li & Stoica, 2008) and its millimeter scale wavelength, radar measurements are noisy, ambiguous, and do not provide elevation. Therefore, radar measurements are often not returned when objects exist or returned when objects do not exist; hence, naively adopting LiDAR methods to radar shows very limited performance on complex scenarios, as in Tables 4 and 5 (CenterPoint (Yin et al., 2021) with radar input). In this paper, we exploit radar data in an adaptive manner to handle its sparsity and ambiguity.

3.2 Radar-assisted View Transformation (RVT)

Image Feature Encoding and Depth Distribution. Given a set of N surrounding images, we use an image backbone (*e.g.*, ResNet (He et al., 2016)) with a feature pyramid network (FPN) (Lin et al., 2017) and obtain 16× downsampled feature map $\mathbf{F}_I$ for each image view. Then, additional convolutional layers further extract image context features $\mathbf{C}_I^{PV} \in \mathbb{R}^{N \times C \times H \times W}$ and depth distribution of each pixel $\mathbf{D}_I \in \mathbb{R}^{N \times D \times H \times W}$ in perspective view, following LSS (Philion & Fidler, 2020):

$$\mathbf{C}_I^{PV} = \text{Conv}(\mathbf{F}_I), \quad \mathbf{D}_I(u, v) = \text{Softmax}(\text{Conv}(\mathbf{F}_I)(u, v)), \tag{3}$$

where (u, v) indicates coordinate in the image plane, and D is the number of depth bins.

Radar Feature Encoding and Radar Occupancy. Unlike previous methods (Philion & Fidler, 2020; Reading et al., 2021; Li et al., 2023b) that directly "lift" image features into BEV using *estimated* depth distribution as Eq. 1, we exploit noisy yet accurate radar measurements for view transformation. Radar points are first projected into each N camera view to find corresponding image pixels while preserving its depth, then voxelized (Lang et al., 2019) into camera frustum view voxels $\mathbf{V}_P^{FV}(d, u, v)$. Note that u, v is pixel unit in the image width and height directions, d is a metric unit in a depth direction. We set $v = 1$ to use pillar-style since radars do not provide reliable elevation measurements. The non-empty radar pillars are encoded into features $\mathbf{F}_P \in \mathbb{R}^{N \times C \times D \times W}$ with PointNet (Qi et al., 2017) and sparse convolution (Yan et al., 2018). Similar to Eq. 3, we extract radar context feature $\mathbf{C}_P^{FV} \in \mathbb{R}^{N \times C \times D \times W}$ and radar occupancy $\mathbf{O}_P \in \mathbb{R}^{N \times 1 \times D \times W}$ in frustum view. Here, convolution is applied to top-view (d, u) coordinate instead of (u, v):

$$\mathbf{C}_P^{FV} = \text{Conv}(\mathbf{F}_P), \quad \mathbf{O}_P(d, u) = \sigma(\text{Conv}(\mathbf{F}_P)(d, u)). \tag{4}$$

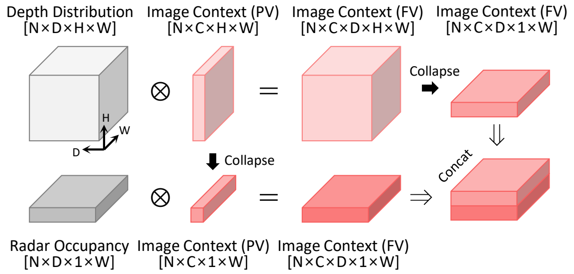
Depth Distribution [N×D×H×W]
Image Context (PV) [N×C×H×W]
Image Context (FV) [N×C×D×H×W]
Image Context (FV) [N×C×D×1×W]
Collapse
H
W
D
Collapse
Concat
Radar Occupancy [N×D×1×W]
Image Context (PV) [N×C×1×W]
Image Context (FV) [N×C×D×1×W]

Figure 3: Radar-assisted View Transformation (RVT). The proposed RVT can benefit from dense but less accurate depth distribution and sparse but accurate radar occupancy to obtain spatially accurate image context features.

Here, a sigmoid is used instead of softmax since radar occupancy is not necessarily one-hot encoded as a depth distribution.

Frustum View Transformation. Given depth distribution $\mathbf{D}_I$ and radar occupancy $\mathbf{O}_P$, the image context feature map $\mathbf{C}_I^{PV}$ is transformed into a camera frustum view $\mathbf{C}_I^{FV} \in \mathbb{R}^{N \times C \times D \times H \times W}$ by the outer product as:

$$\mathbf{C}_I^{FV} = \text{Conv}[\mathbf{C}_I^{PV} \otimes \mathbf{D}_I; \mathbf{C}_I^{PV} \otimes \mathbf{O}_P], \tag{5}$$

where $[\cdot;\cdot]$ denotes the concatenation operating along the channel dimension. Due to the absence of height dimension in radar and for saving memory, we collapse the image context feature by summation along the height axis, as illustrated in Fig. 3.

Bird's-Eye-View Transformation. Finally, camera and radar context feature maps in N camera frustum views $\mathbf{F} = \{\mathbf{C}_I, \mathbf{C}_P \in \mathbb{R}^{N \times C \times D \times H \times W}\}$ are transformed into a single BEV space $\mathbb{R}^{C \times 1 \times X \times Y}$ by view transformation module $\mathcal{M}$:

$$\mathbf{F}^{BEV} = \mathcal{M}(\{\mathbf{F}_i^{FV}\}_{i=1}^{N}). \tag{6}$$

Specifically, we adopt CUDA-enabled Voxel Pooling (Li et al., 2023a) implementation and modify it to aggregate features within each BEV grid using average pooling instead of summation. It helps the network to predict a more consistent BEV feature map regardless of the distance to the ego vehicle since a closer BEV grid is associated with a more frustum grid due to the perspective projection.

3.3 Multi-modal Feature Aggregation (MFA)

Motivation. Combining complementary multi-modal information while avoiding the drawbacks of each is especially crucial in camera radar fusion, as claimed in Sec. 3.1. Image feature has rich semantic cues but is inherently spatially inaccurate; on the other hand, radar feature is spatially accurate, but contextual information is very limited and noisy. Naive approaches are channel-wise concatenation (Liu et al., 2022b) or summation (Li et al., 2022a), but these cannot handle neither spatial misalignment nor ambiguity between two modalities, thus less effective, as can be seen in Table 3. To have the best of both worlds, the key motivation of our fusion is to leverage multi-modal features in an adaptive manner, using an attention mechanism (Vaswani et al., 2017).

Multi-modal Deformable Cross Attention. Cross attention (Vaswani et al., 2017) is inherently suitable for multi-modal fusion, but the computation cost is quadratic to input sequence length $\mathcal{O}(N^2)$, where $N = HW$ and H, W denote the height and width of the BEV feature map. If we assume perception range $R = H/2 = W/2$, computation complexity becomes biquadratic $\mathcal{O}(16R^4)$ to perception range, which is not scalable for a long-range perception; Thus we develop the fusion method based on deformable attention (Zhu et al., 2021), which is of linear complexity with the input size $\mathcal{O}(2N + NK)$, where K is the total number of the sampled key ($K \ll N = HW$).

Given BEV context feature maps $\text{x}_m = \{\mathbf{C}_I, \mathbf{C}_P\}$, we project x_m into C dimensional query feature after concatenation as $\text{z}_q = \boldsymbol{W}_z[\text{LN}(\mathbf{C}_I); \text{LN}(\mathbf{C}_P)]$, where $\boldsymbol{W}_z \in \mathbb{R}^{2C \times C}$ is a linear projection and LN is layer norm. After, the feature map is aggregated by multi-modal deformable cross attention as

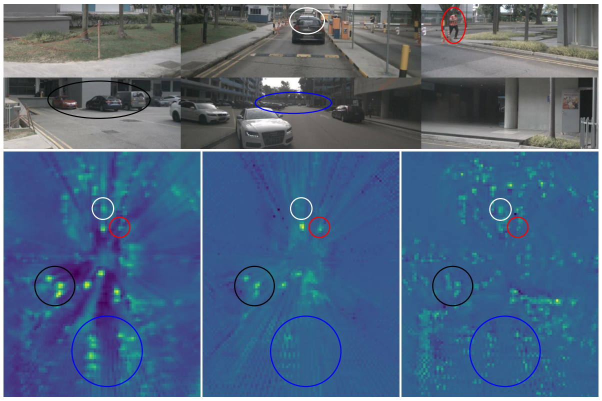
(a) Fusion (b) Image (c) Radar

Figure 4: In image (b), a vehicle heavily occluded (white) or hardly visible at a long distance (blue) is not detected. In radar (c), clutters from the wall (black) or pedestrian with row RCS (red) lead to failure. Our MFA (a) generates a more reliable BEV feature map by fusion. Note that BEV feature maps are cropped for better visualization.

$$\text{MDCA}(\text{z}_q, p_q, \text{x}_m) = \sum_{h}^{H} \boldsymbol{W}_h \left[\sum_{m}^{M} \sum_{k}^{K} A_{hmqk} \cdot \boldsymbol{W}'_{hm} \text{x}_m(\phi_m(p_q + \Delta p_{hmqk})) \right], \tag{7}$$

where h, m, k indexes the attention head, modality, and sampling point. To better exploit multi-modal information, we separately apply attention weights A_{hmqk} and sampling offset Δp_{hmqk} to multi-modal feature maps x_m. By doing so, the feature aggregation module can adaptively benefit from image and radar as shown in Fig. 4. We refer the reader to Appendix for details of the notation.

Sparse Aggregation. Although MDCA has linear complexity with respect to BEV size, it still can be a bottleneck when the perception range becomes large. Inspired by (Yao et al., 2021; Roh et al., 2022), we propose a method to further reduce the number of input queries from $N = HW$ to $N = N_k \ll HW$ by using features with top-k confidence.4 Given BEV depth distribution $\mathbf{D}_I$ and radar occupancy $\mathbf{O}_P$, N_k features $\mathrm{z}_q^{N_k} \in \mathbb{R}^{C \times N_k}$ are selected from input queries $\mathrm{z}_q \in \mathbb{R}^{C \times HW}$ using a probability of $\max(\mathbf{D}_I, \mathbf{O}_P)$. The complexity of the proposed sparse aggregation is now independent of perception range, which is more efficient for long-range perception.

3.4 Training Objectives and Task Heads

We train the depth distribution network with a depth map obtained by projecting LiDAR points into the image view, following BEVDepth (Li et al., 2023b). We follow CenterPoint (Yin et al., 2021) to predict the center heatmap with anchor-free and multi-group head (Zhu et al., 2019). For training sparse aggregation, we filter LiDAR points outside of 3D bounding box when obtaining a ground truth depth map and replace the softmax to sigmoid in Eq. 3; thereby, only feature grids containing foreground objects can have a high probability.

4 Experiments

4.1 Experimental Settings

Dataset and Metrics. We conduct experiments on 3D object detection task on nuScenes (Caesar et al., 2020), which is the only dataset providing radar point cloud at scale. We use official metrics: mAP (Everingham et al., 2010) and NDS (Caesar et al., 2020) and we refer the reader to nuScenes (Caesar et al., 2020) for details of metrics.

Implementation Details. For the camera stream, we adopt BEVDepth (Li et al., 2023b) as a baseline with several modifications. We reduce the number of depth estimation layers and eliminate the depth refinement module, which increases the inference speed without a significant performance drop. For radar, we accumulate six previous radar sweeps and use normalized RCS and Doppler speed as features following GRIF Net (Kim et al., 2020b). Unless otherwise specified, we follow standard practices (Huang et al., 2021; Li et al., 2023b) for implementation and training details. We accumulate four BEV feature maps with an interval of 1 second, similar to BEVFormer (Li et al., 2022c). The full experimental settings are provided in Appendix.

4.2 Main Results

For a comparison with previous state-of-the-art methods, we train and evaluate our model on 3D detection task and report `val` set results in Table 1. Under various input sizes and backbone settings, our CRN achieves first place among all camera-only and camera-radar methods with much faster FPS (Sec. 4.4 for inference time analysis). We emphasize that CRN with a small input size and backbone (256×704 and R50) already achieves a competitive performance with a large input size and backbone (BEVDepth (Li et al., 2023b) and SOLOFusion (Park et al., 2022) with 512×1408 and R101) in terms of mAP while running an order of magnitude faster, showing the effectiveness of using radar over camera-only methods. Ours also outperforms the LiDAR method CenterPoint-P (Yin et al., 2021), demonstrating the potential of cost-effective camera and radar to replace LiDAR for autonomous driving. Qualitative results are provided in Appendix.

4.3 Ablation Studies

We conduct ablation studies on `val` set with 3D detection task. Unless otherwise specified, models use two frames of 256×704 image, R50 backbone, and are trained for 24 epochs without CBGS (Zhu et al., 2019). For thorough comparison, we additionally build three baseline detectors for camera – BEVDepth (Li et al., 2023b), point – CenterPoint (Yin et al., 2021), and camera-point

Table 1: **3D Object Detection** on nuScenes `val` set. 'L', 'C', and 'R' represent LiDAR, camera, and radar, respectively. *: results from MMDetection3D (Chen et al., 2019). †: trained with CBGS.

	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS
CenterPoint-P$^{\dagger *}$ (Yin et al., 2021)	L	Pillars	-	59.8	49.4	0.320	0.262	0.377	0.334	0.198	-
CenterPoint-V$^{\dagger *}$ (Yin et al., 2021)	L	Voxel	-	65.3	56.9	0.285	0.253	0.323	0.272	0.186	-
BEVDet† (Huang et al., 2021)	C	R50	256 × 704	39.2	31.2	0.691	0.272	0.523	0.909	0.247	15.6
CenterFusion† (Nabati & Qi, 2021)	C+R	DLA34	448 × 800	45.3	33.2	0.649	**0.263**	0.535	0.540	**0.142**	-
BEVDepth† (Li et al., 2023b)	C	R50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198	11.6
RCBEV4d† (Zhou et al., 2023)	C+R	Swin-T	256 × 704	49.7	38.1	0.526	0.272	0.445	0.465	0.185	-
CRAFT† (Kim et al., 2023)	C+R	DLA34	448 × 800	51.7	41.1	0.494	0.276	0.454	0.486	0.176	4.1
SOLOFusion† (Park et al., 2022)	C	R50	256 × 704	53.4	42.7	0.567	0.274	**0.411**	**0.252**	0.188	11.4
CRN	C+R	R50	256 × 704	**56.0**	**48.1**	**0.474**	0.271	0.541	0.328	0.188	**20.4**
FCOS3D (Wang et al., 2021b)	C	R101	900 × 1600	41.5	34.3	0.725	**0.263**	0.422	1.292	**0.153**	1.7
DETR3D† (Wang et al., 2022)	C	R101	900 × 1600	43.4	34.9	0.716	0.268	0.379	0.842	0.200	-
PETR† (Liu et al., 2022a)	C	R101	900 × 1600	44.2	37.0	0.711	0.267	0.383	0.865	0.201	1.7
BEVFormer (Li et al., 2022c)	C	R101	900 × 1600	51.7	41.6	0.673	0.274	0.372	0.394	0.198	1.7
PolarFormer-T (Jiang et al., 2023)	C	R101	900 × 1600	52.8	43.2	0.648	0.270	**0.348**	0.409	0.201	1.7
BEVDepth† (Li et al., 2023b)	C	R101	512 × 1408	53.5	41.2	0.565	0.266	0.358	0.331	0.190	5.0
SOLOFusion (Park et al., 2022)	C	R101	512 × 1408	54.4	47.2	0.518	0.275	0.604	0.310	0.210	-
SOLOFusion† (Park et al., 2022)	C	R101	512 × 1408	58.2	48.3	0.503	0.264	0.381	**0.246**	0.207	-
CRN	C+R	R101	512 × 1408	**58.3**	**51.5**	**0.463**	0.268	0.447	0.370	0.192	**7.2**

Table 2: Ablation of view transformation methods. LiDAR and radar are used only for transformation and not used for feature aggregation.

Input	RVT	*All*			*Car*
		NDS	mAP	mATE	mAP
Depth	✗	43.9	33.2	0.716	50.4
Radar	✗	33.6	24.3	0.706	44.7
Depth+Radar	✓	52.1	44.8	0.521	70.5
Depth+LiDAR	✓	57.0	51.6	0.419	76.2

Table 3: Ablation of feature aggregation methods. Note that MFA with RVT is our full model.

	Input	*All*			*Car*
		NDS	mAP	mATE	mAP
CenterPoint	L	52.8	41.2	0.406	73.9
BEVFusion	C+R	51.9	42.4	0.536	68.4
+ deeper conv	C+R	51.9	42.8	0.532	69.0
+ RVT	C+R	52.7	44.3	0.517	70.6
MFA	C+R	53.4	44.5	0.507	70.3
+ RVT	C+R	**53.9**	**45.2**	**0.501**	**71.6**

– BEVFusion (Liu et al., 2022b). Details of baselines and additional ablation studies are provided in Appendix.

View Transformation. In Table 2, we study how the radar-assisted feature transformation affects performance. View transformation solely relying on estimated depth suffers from inaccurate localization due to the inherent low accuracy of depth distribution. If we naively replace depth distribution to radar (1 if radar point exists inside the voxel, 0 else), performance is severely degraded. This is because image features in perspective view cannot be properly transformed due to the ambiguity and sparsity of radar. With the proposed RVT, the model can benefit from both dense depth and sparse range measurement to significantly improve performance (+8.2% NDS, +11.6% mAP) over depth-only transformation. Moreover, we find consistent performance improvement on LiDAR input, showing the effectiveness of RVT.

Feature Aggregation. Table 3 shows the comparison between different feature aggregation methods. BEVFusion (Liu et al., 2022b) fuses multi-modal feature maps in BEV using a single convolutional layer, which is not adaptive and has a small receptive field (3 × 3). Simply adding two additional convolutional layers for fusion, which provides a larger receptive field (7 × 7) and bigger capacity, does not improve the performance much. On the other hand, using only MFA already outperforms deeper BEVFusion with RVT, showing the effectiveness of the proposed multi-modal deformable cross attention. We find that the performance gain of RVT is less significant on MFA than BEVFusion since MFA is already capable of handling spatial misalignment between multi-modal features without RVT.

4.4 Analysis

Scaling Up Perception Range. In Table 4, we increase the perception range of BEV grids from 51.2m to 102.4m and also increase the evaluation range twice correspondingly (see Appendix for details). Although CenterPoint (Yin et al., 2021) uses 10 LiDAR sweeps, points become extremely sparse as the range increases, and thus performance is significantly degraded at far distances. On

Table 4: Analysis over various perception range. Suffix -S denotes sparse aggregation and we use 256×704 and R50 for all camera streams.

	Input	*Car* mAP				FPS
		[0,100)	[0,30)	[30,60)	[60,100)	
CenterPoint	L	54.2	**84.3**	35.8	4.8	6.3
BEVDepth	C	34.1	65.4	13.7	0.2	13.0
CenterPoint	R	20.3	36.6	11.6	2.9	**30.7**
CRN	C+R	**56.9**	82.6	**42.6**	**7.0**	11.5
CRN-S	C+R	54.0	79.2	39.8	6.2	14.0

Table 5: Analysis of robustness using *Car* class mAP. Six view drops denote the single modality is entirely off.

	Input	Drop	# of view drops			
			0	1	3	6
BEVDepth	C	C	49.4	41.1	24.2	0
CenterPoint	R	R	30.6	25.3	14.9	0
BEVFusion	C+R	C	63.9	58.5	45.7	**14.3**
		R		59.9	50.9	34.4
CRN	C+R	C	**68.8** (+4.9)	**62.4**(+3.9)	**48.9**(+3.2)	12.8(-1.5)
		R		**64.3**(+4.4)	**57.0**(+6.1)	**43.8**(+9.4)

the other hand, CRN outperforms LiDAR especially at farther than $30m$ range with a much faster FPS, showing the effectiveness of camera and radar for a long range perception. Moreover, CRN with sparse aggregation further improves the inference speed while preserving the comparable performance.

Robustness. To systematically analyze the robustness of sensor failure cases, we randomly drop image and radar inputs in Table 5. For fair comparisons, we use *single* frame input and fix the seed to ensure the same views can be dropped over experiments. We also train both fusion methods with data-level augmentation (Chen et al., 2022). CRN not only outperforms BEVFusion when all modalities are available but maintains higher mAP on sensor failure cases. Considering that ours uses radar points at multiple stages (RVT and MFA), each proposed module is trained to be robust to sparse and ambiguous radar points. Especially when radar input is entirely off, BEVFusion suffers from a performance drop over BEVDepth (-15.0%), while CRN still keeps the competitive performance (-5.6%). This advantage comes from our attention module, which can adaptively choose modalities to use.

Inference Time. We analyze the inference time of each proposed component in Fig. 5. For all analyses, we assume that the BEV feature map of the previous frame $T-1$ can be stored and accessed at the current frame T since ours does not use temporal information (*e.g.*, (Li et al., 2023a; Park et al., 2022)) when obtaining the BEV feature map. It means that using a multi-frame only increases the latency of the BEV head. Ours requires negligible additional computation for point encoder and fusion modules, but the performance gain over additional latency is substantial (+14.9ms for +12.4 NDS in 256x704 and R50 setting). Moreover, ours with small input can outperform camera-only with larger input in terms of both latency and performance. We expect that inference optimization methods (*e.g.*, TensorRT) can further reduce the latency of large model for long perception range setting to match the real-time.

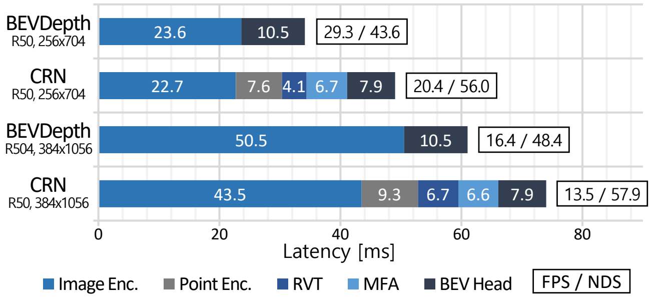

Figure 5: Inference time analysis of proposed components. All models are trained without CBGS (Zhu et al., 2019) and latency numbers are measured with batch size 1 and GPU warm-up.

5 Conclusion

We present CRN, a novel camera-radar fusion method for accurate, robust, and efficient 3D perception. Our method effectively overcomes the limitation of each modality and efficiently fuses multi-modal information to generate contextually rich and spatially accurate BEV features. CRN is also suitable for long-range perception in real-time and achieves state-of-the-art performance on nuScenes benchmarks. We hope that CRN will inspire future research on camera-radar fusion for 3D perception.

REFERENCES

Continental ARS 408-21 Datasheet. https://conti-engineering.com/components/ars-408/. Accessed: 2023-03-01.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1090–1099, 2022.

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)*, pp. 11621–11631, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

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)*, 2022.

Hyunggi Cho, Young-Woo Seo, BVK Vijaya Kumar, and Ragunathan Raj Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1836–1843, 2014.

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

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)*, pp. 248–255, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (Int. J. Comput. Vis.)*, 88(2):303–338, 2010.

Daniel Göhring, Miao Wang, Michael Schnürmacher, and Tinosch Ganjineh. Radar/lidar sensor fusion for car-following on highways. In *Proceedings of the IEEE International Conference on Automation, Robotics and Applications (ICARA)*, pp. 407–412, 2011.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D Packing for Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2485–2494, 2020.

Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? *arXiv preprint arXiv:2206.07959*, 2022.

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)*, pp. 770–778, 2016.

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)*, pp. 15273–15282, 2021.

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.

Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Don H Johnson and Dan E Dudgeon. *Array signal processing: concepts and techniques*. Simon & Schuster, Inc., 1992.

Jinhyeong Kim, Youngseok Kim, and Dongsuk Kum. Low-level sensor fusion network for 3d vehicle detection using radar range-azimuth heatmap and monocular image. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 388–402, 2020a.

Youngseok Kim, Jun Won Choi, and Dongsuk Kum. GRIF Net: Gated region of interest fusion network for robust 3D object detection from radar point cloud and monocular image. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10857–10864, 2020b.

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.

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)*, pp. 12697–12705, 2019.

Jian Li and Petre Stoica. *MIMO radar signal processing*. John Wiley & Sons, 2008.

Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17182–17191, 2022b.

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)*, 2023a.

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)*, 2023b.

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)*, pp. 1–18, 2022c.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10233–10240, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017.

Yier Lin, Julien Le Kernec, Shufan Yang, Francesco Fioranelli, Olivier Romain, and Zhiqin Zhao. Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests. *IEEE Sensors Journal*, 18(23):9669–9681, 2018.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 531—548, 2022a.

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 *arXiv preprint arXiv:2205.13542*, 2022b.

Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12507–12516, 2021.

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

Jiachen Lu, Zheyuan Zhou, Xiatian Zhu, Hang Xu, and Li Zhang. Learning ego 3d representation as ray tracing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 129–144, 2022.

Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 924–932, 2019.

Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *Proceedings of the European Radar Conference (EuRAD)*, pp. 129–132, 2019.

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)*, pp. 1527–1536, 2021.

Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10386–10393. IEEE, 2020.

Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3142–3152, 2021.

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 *arXiv preprint arXiv:2210.02443*, 2022.

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)*, pp. 194–210, 2020.

Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 918–927, 2018.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5105–5114, 2017.

Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8555–8564, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, 2015.

Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11138–11147, 2020.

Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9200–9206, 2022.

Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1991–1999, 2019.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9627–9636, 2019.

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)*, pp. 6000–6010, 2017.

Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4604–4612, 2020.

Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 1475–1485, 2021a.

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 Workshops (ICCVW)*, pp. 913–922, 2021b.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 180–191, 2022.

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

Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

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)*, pp. 11784–11793, 2021.

Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2020.

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

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)*, pp. 13760–13769, 2022.

Taohua Zhou, Junjie Chen, Yining Shi, Kun Jiang, Mengmeng Yang, and Diange Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. *IEEE Transactions on Intelligent Vehicles (IEEE Trans. Intell. Veh.)*, 2023.

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

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.

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.

Appendix

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 (Sec. C) for experiments in the main paper. We further provide additional experimental results (Sec. D) and qualitative results (Sec. E).

B Multi-modal Deformable Cross Attention

We adapt the deformable attention (Zhu et al., 2021) and extend it for multi-modal feature maps, denoted as multi-modal deformable cross attention.

Given an input queries z_q and multi-modal feature maps $\mathrm{x}_m = \{\mathbf{C}_I, \mathbf{C}_P \in \mathbb{R}^{C\times H\times W}\}$, 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 $\{\mathbf{C}_I, \mathbf{C}_P\}$, 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 (Vaswani et al., 2017). 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

In this section, we provide the experimental settings for 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 (Yan et al., 2018), which concatenates output feature maps at stride 16. `nn.Conv2d` and `nn.ConvTranspose2d` are used for downsampling and upsampling in SECONDFPN. 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 $D = 112$.

Table 6: Training settings for the main results.

configs	ResNet-50	ResNet-101
optimizer	AdamW	
weight decay	1e-4	
base lr	2e-4	1e-4
backbone lr	2e-4	5e-5
batch size	64	32
training epochs	24	
lr schedule	step decay	
gradient clip	5	

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. After, 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 MMCV (Contributors, 2018) `multi_scale_deform_attn` implementation for deformable cross attention in Multi-modal Feature Aggregation (MFA). Specifically, we use 6 layers of MFA, 8 attention head, and 4 sampling points for MFA in our experiments.

Following standard practices in monocular 3D object detection (Huang et al., 2021; Li et al., 2023b), 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 BEV feature map has 128×128 size.

C.2 Training Settings

All models are trained for 24 epochs with AdamW (Loshchilov & Hutter, 2019) optimizer in an end-to-end manner. Image backbones are pre-trained on ImageNet (Deng et al., 2009). We provide ResNet (He et al., 2016) 50 and 101 training settings used for our main results in Table 6.

For image and radar data augmentation (in perspective view), we use resize, crop, and horizontal flipping augmentation following standard practices (Huang et al., 2021; Li et al., 2023b). 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) (Yan et al., 2018) is not used in our experiments, and we leave GT-AUG for a multi-modal setting (Chen et al., 2022; Zhang et al., 2020) as future work.

C.3 Baselines for Ablation Studies

We conduct three baselines BEVDepth (Li et al., 2023b), CenterPoint (Yin et al., 2021), and BEVFusion (Liu et al., 2022b) for camera-only, point-only, and camera-point fusion detectors. For BEVDepth, we use the official code[1] without class-balanced grouping and sampling (CBGS) (Zhu et al., 2019) and exponential moving average (EMA). For CenterPoint, we use MMDetection[2] implementation using PointPillar (Lang et al., 2019) backbone with $(0.2, 0.2, 8)m$ pillar size. Unlike the official implementation, CBGS (Zhu et al., 2019) and ground-truth sampling augmentation (Yan et al., 2018) 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

[1] https://github.com/Megvii-BaseDetection/BEVDepth

[2] https://github.com/open-mmlab/mmdetection3d

yield better performance since our implementation uses BEVDepth for the camera stream, while the original BEVFusion uses LSS (Philion & Fidler, 2020).

C.4 Details of Long Range Perception

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 bin becomes $D = 224$. 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 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 (Caesar et al., 2020) 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 for training and evaluation, but we disable this filtering for thorough analysis. Thus, some moving objects are visible on the image but cannot 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 harm 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 (Yin et al., 2021) with radar input is much faster than LiDAR input, assuming that the sparsity of radar points can highly benefit from voxelization and sparse convolution (Yan et al., 2018).

D Additional Experimental Results

D.1 Design Decisions

We study architecture parameter decisions that affect the performance of CRN to provide insights on the proposed sensor fusion framework. All experiments are conducted on nuScenes `val` set.

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 (Li et al., 2022c; 2023b). Compared to temporal stereo methods (Li et al., 2023a; Park et al., 2022), 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 7, using multiple temporal frames significantly improves mAP, mATE, and mAVE. Corresponding to results on recent approaches using temporal BEV feature maps (Park et al., 2022), 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 8, 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

Table 7: Ablation of temporal frames.

# 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

Table 8: Ablation of sparse aggregation.

# 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

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 are more significant than AP, assuming that the classification network can maintain its performance but the regression network suffers from sparsely spread BEV features to regress objects' attributes.

D.2 Weather and Lighting Analysis

We analyze the performance under different weather and lightning conditions in Table 9. Note that R101 backbone with 512×1408 input is used for BEVDepth and ours for comparable comparisons with LiDAR methods. Sensor noises of LiDAR in rainy conditions or poor illumination of camera at night make object detection challenging for LiDAR-only or camera-only methods. Thanks to fusion with radar, ours shows consistent performance improvement of more than 10 mAP over the camera-only method, demonstrating the effectiveness and robustness of camera and radar sensors in all weather conditions.

Table 9: Analysis over different lighting and weather conditions using mAP metric. CenterPoint (Yin et al., 2021) results are from BEVFusion (Liu et al., 2022b), and BEVDepth results are reproduced by us.

	Input	Sunny	Rainy	Day	Night
CenterPoint (Yin et al., 2021)	L	62.9	59.2	62.8	35.4
RCBEV (Zhou et al., 2023)	C+R	36.1	38.5	37.1	15.5
BEVDepth (Li et al., 2023b)	C	39.0	39.0	39.3	16.8
CRN	C+R	51.6(+12.6)	54.3(+15.3)	52.0(+12.7)	28.3(+11.5)

D.3 Per-Class Analysis

In Table 10, we compare the performance improvement of camera-radar methods over camera-only baselines. For fair comparisons, all models are trained with CBGS (Zhu et al., 2019), and we report 256×704 and R50 models for BEVDepth and CRN. Corresponds to results on CRAFT (Kim et al., 2023), 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 on ours is much more significant than other fusion methods. Considering the performance of camera baselines are similar, it demonstrates that the design of fusion methods greatly affects the performance.

E Qualitative Results

We visualize the 3D detection results of 256×704 and R50 model in Fig. 6. As can be seen, 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 succesfully 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

Table 10: 'C.V.', 'Ped.', 'M.C.', and 'T.C.' denote construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. CenterNet (Zhou et al., 2019), CRAFT-I (Kim et al., 2023), and BEVDepth (Li et al., 2023b) are camera baselines of CenterFusion (Nabati & Qi, 2021), CRAFT (Kim et al., 2023), and CRN. CenterPoint and BEVDepth results are from MMDetection3D and their official code.

Method	Input	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	Bicycle	T.C.	Barrier	mAP
CenterPoint-P	L	83.9	49.5	61.9	34.1	12.3	76.9	44.1	18.0	54.0	59.1	49.4
CenterNet	C	48.4	23.1	34.0	13.1	3.5	37.7	24.9	23.4	55.0	45.6	30.6
CenterFusion	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	C	52.4	25.7	30.0	15.8	5.4	39.3	28.6	29.8	57.5	47.8	33.2
CRAFT	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	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	74.7(**+19.4**)	42.8(**+17.6**)	50.3(**+12.5**)	22.2(+5.9)	12.6(+5.0)	53.9(**+17.8**)	48.5(**+16.6**)	41.5(**+12.9**)	61.7(+8.1)	63.4(+7.5)	47.2(**+12.4**)

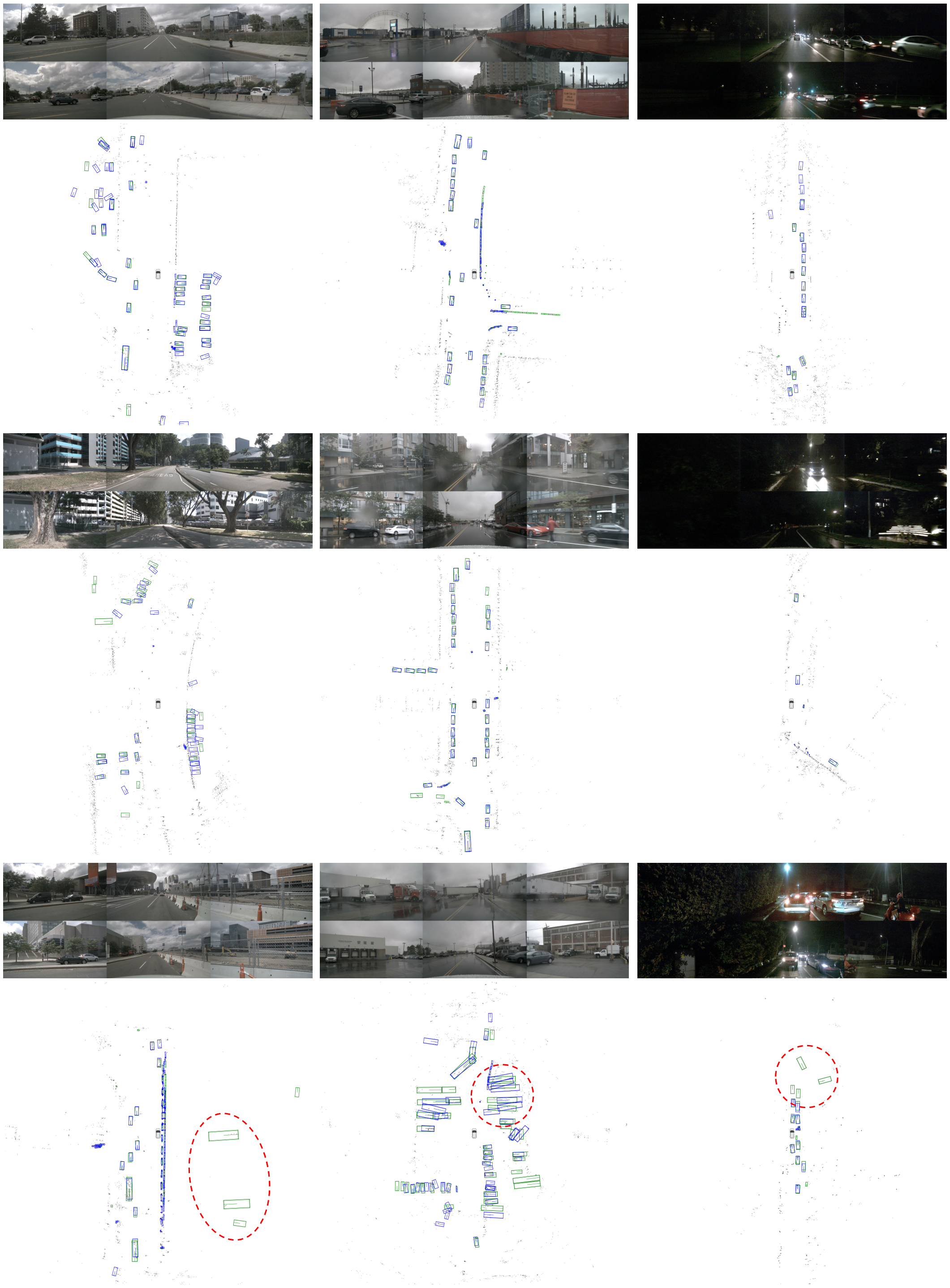

Figure 6: Qualitative results on nuScenes `val` set: from left to right, Day, Rainy, and Night scenarios. We show the failure cases and highlight them with red circles on the bottom row. Green boxes are ground truths, blue boxes are our prediction results, and black dots are radar points. The perception range is set to $100m \times 100m$ and best viewed in color with zoom in.

rare classes and do not without radar points (*e.g.*, construction vehicles behind wire mesh or trailers heavily occluded).