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

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

Abstract

Autonomous driving requires an accurate and fast 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 image and radar feature maps in BEV using multi-modal deformable attention designed to tackle the spatial misalignment between inputs. CRN with real-time setting operates at 20 FPS while achieving comparable performance to LiDAR detectors on nuScenes, and even outperforms at a far distance on 100m *setting. Moreover, CRN with offline setting yields 62.4% NDS, 57.5% mAP on nuScenes test set and ranks first among all camera and camera-radar 3D object detectors.*

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 [80], BEV segmentation [82], HD map generation [64], trajectory prediction [17], and motion planning [52]). Another important ingredient for deploying 3D perception to the real world is to build a system that relies less on LiDAR disadvantaged from high-cost, high-maintenance, and low-reliability. 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 benefited from radar.

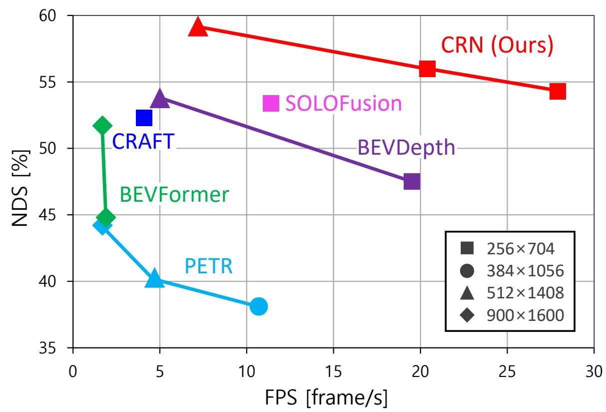


Figure 1. FPS vs. accuracy on nuScenes `val` set. We show that fusing radar can significantly boost camera-only method with marginal computational cost. CRN outperforms all methods with much faster speed. See Table 1 and Fig. 6 for more details.

Recently, camera-based 3D perception in BEV [19, 52, 57] 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, localizing the 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 also in high-reliability, long-range perception (up to 200m for typical automotive radar [8]), 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× fewer than LiDAR points per single frame in nuScenes [2]), 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 [7, 13] fail to fully exploit the complementary information, thus having limited performance and operating environment. Despite the huge potential of learning-based fusion, only a few studies [23, 24, 48] 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, transformation and fusion 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 image features in perspective view into BEV not solely relying on estimated depth 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)* layers 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 on various tasks while maintaining high efficiency, as shown in Fig 1.

The main contributions of our works are three-fold:

- **Accuracy.** CRN achieves LiDAR-level performance using camera and radar on 3D object detection, tracking, and BEV segmentation tasks.
- **Robustness.** CRN maintains robust performance even if one of the single sensor inputs is entirely unavailable, which allows the fault-tolerant system.
- **Efficiency.** 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 [70, 84] on perspective view images, early approaches extend 2D detector to 3D detector by additionally estimating the distance to objects [65, 74], then transforming object center. DD3D [50] improves detection performance by pre-training depth estimation task on depth dataset [14]. 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 [32, 51, 52, 57] explicitly estimate the depth distribution of each image feature on PV and transform them into BEV. BEVDepth [32] empirically shows that training depth distribution with auxiliary pixel-wise depth supervision improves the performance, which corresponds to the results of DD3D [50]. Learning-based methods [20, 33, 45, 82] implicitly model the mapping function from PV to BEV using multi-layer perceptron (MLP) [29, 58] or attention [33, 60].

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 [32], segmentation [82], and prediction [52]. However, camera-only methods have limited localization accuracy due to the absence of distance information in image and are sensitive to lighting or weather conditions. Moreover, achieving high performance only using a camera requires large image input and backbone, which is slow and not applicable for real-time applications.

Point-based 3D Perception. LiDAR is the most common and favorable sensor for autonomous driving, while radar point cloud has not yet been thoroughly investigated. LiDAR-only 3D detectors extract features (*e.g.*, PointNet [55, 56]) given irregular and unordered point sets and predict 3D objects on point- [62] or voxelized- [26] feature. Some approaches further utilize point and voxel features together [61], use range view as additional features [75], or filter background points [67].

Although similar data representation of radar point cloud [2, 47] to LiDAR, radar point-based 3D perception is considerably less investigated. Several works [53, 66, 77] examine the radar points for free space detection, but only a few studies [68, 71] attempt 3D object detection in autonomous driving. Radar point-based detection methods adapt PointPillars [26] with graph neural network [63] or KPConv [69] focusing on extracting better local features. However, mostly due to many clutter points and lack of contextual information on radar, the performance of radar-only methods lags significantly behind compared to LiDAR. Considering the high potential of radar having robust measurements regardless of weather conditions and perception range, fusing radar with a camera is promising to supplement the insufficient semantic information.

Camera-Point 3D Perception. Fusing complementary information of camera image and range measurement is a

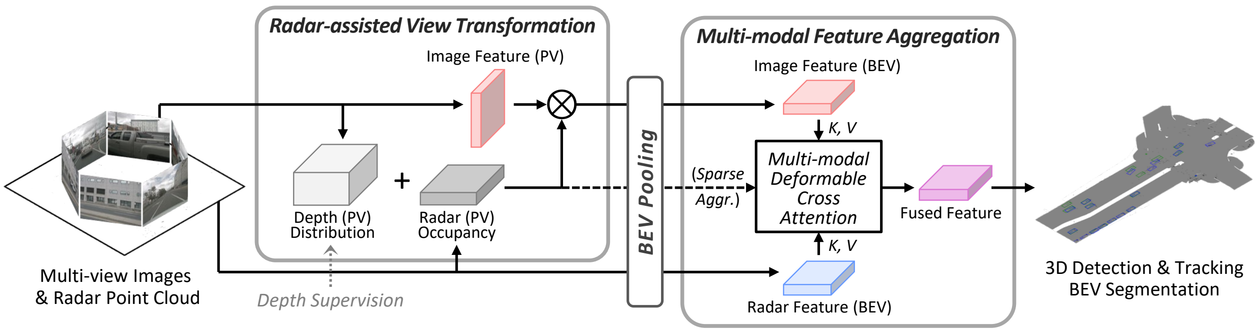


Figure 2. The overall architecture of the proposed Camera Radar Net. Given multi-view images and radar points, modality-specific backbones extract features in each view. First, image context features in perspective view are transformed into a bird's-eye-view with the help of radar measurements by Radar-assisted 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.

promising and active research topic. However, the view discrepancy between two sensors is regarded as a bottleneck for multi-modal fusion. A line of approach handles discrepancy by projecting 3D information to a 2D image (*e.g.*, points [6, 73, 78], proposals [1, 23, 25], or prediction results [49]) and gathering information around the projected region. Some camera-radar fusion methods [35, 43] attempt to improve depth estimation by projecting radar points to the image.

On the other hand, another line of work lifts 2D image information into 3D. Early studies in 3D detection [24, 48, 54] detect 2D or 2.5D object proposals and 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 concatenation [42] or summation [31], 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 [10, 31, 34, 81], BEV segmentation [42, 83], or HD map generation [9, 29]. However, despite the 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. Preliminaries

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 [32, 52, 57] 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 [32, 42]) and $\otimes$ denotes outer product. Meanwhile, learning-based approaches [33, 82] 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) [58] or cross-attention [33]), 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. We aim to explicitly improve the transformation process using radar measurement.

Radar Characteristics. Radar can have various representations (*e.g.*, 2-D FFT [39], 3D Tensor [22, 46], point cloud [2, 47]). Radar point cloud has a similar representation to LiDAR, but their characteristics are different in terms of resolution and accuracy [8]. Moreover, due to the nature of the operating mechanism of radar [21, 28] 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 7 and 8 (CenterPoint [80] with radar input). We exploit radar in an adaptive manner to handle its sparsity and ambiguity.

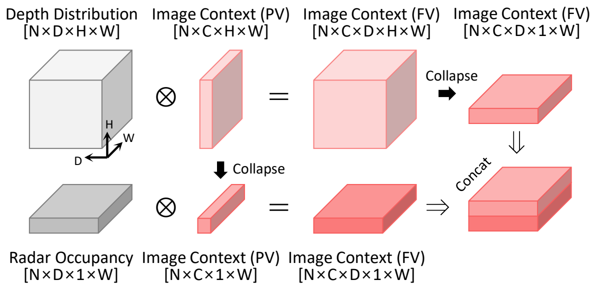


Figure 3. Radar-assisted View Transformation (RVT). The proposed RVT can benefit from dense but less accurate depth distribution and sparse but accurate radar occupancy to obtain spatially accurate image context features.

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 [16], ConvNeXt [41]) with a feature pyramid network (FPN) [36] and obtain $16\times$ 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 [52]:

$$\begin{aligned} \mathbf{C}_I^{PV} &= \mathrm{Conv}(\mathbf{F}_I) \\ \mathbf{D}_I(u,v) &= \mathrm{Softmax}(\mathrm{Conv}(\mathbf{F}_I)(u,v)), \end{aligned} \tag{3}$$

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 [32, 52, 57] 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 onto each N camera view to find corresponding image pixels while preserving its depth, then voxelized [26] into camera frustum view voxels $\mathbf{V}_R^{FV}(d, u, v)$. Note that u, v are pixel units in the image width and height directions, while 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 $\mathbf{F}_R \in \mathbb{R}^{N\times C\times D\times W}$ with PointNet [56] and sparse convolution [79]. Similar to Eq. 3, we extract radar context feature $\mathbf{C}_R^{FV} \in \mathbb{R}^{N\times C\times D\times W}$ and radar occupancy $\mathbf{O}_R \in \mathbb{R}^{N\times 1\times D\times W}$ in frustum view. Here, convolution is applied to top-view (d, u) coordinate instead of (u, v):

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

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 $\mathbf{O}_R$, the image context feature map $\mathbf{C}_I^{PV}$ is transformed into a camera frustum view $\mathbf{C}_I^{FV} \in \mathbb{R}^{N\times C\times D\times H\times W}$ as:

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

where $[\cdot;\cdot]$ denotes the concatenation operating along the channel dimension and $\otimes$ is the outer product. 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 $\mathbf{F}^{FV} = \{\mathbf{C}_I^{FV}, \mathbf{C}_R^{FV} \in \mathbb{R}^{N\times C\times D\times H\times W}\}$ are transformed into a single BEV space $\mathbb{R}^{C\times 1\times X\times Y}$ by view transformation module $\mathcal{M}$:

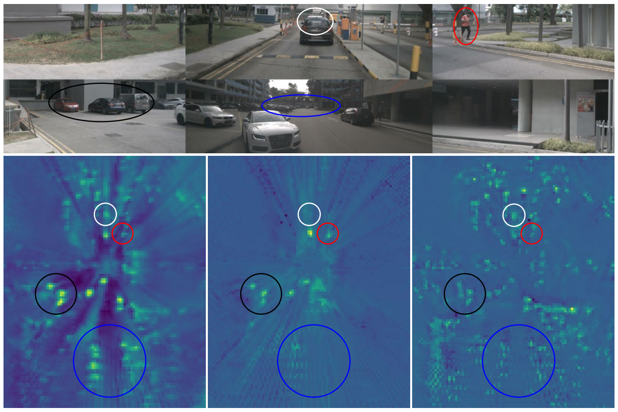
$$\mathbf{F}^{BEV} = \mathcal{M}(\{\mathbf{F}_i^{FV}\}_{i=1}^N). \tag{6}$$

Specifically, we adopt CUDA-enabled Voxel Pooling [30] 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 their spatial position is inherently inaccurate; on the other hand, radar feature is spatially accurate, but contextual information is insufficient and noisy. Naive approaches are channel-wise concatenation [42] or summation [31], but these cannot handle neither spatial misalignment nor ambiguity between two modalities, thus less effective, as can be seen in Table 6. 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 [72].

Multi-modal Deformable Cross Attention (MDCA). Cross attention [72] is inherently suitable for multi-modal fusion. However, the computation cost is quadratic to input sequence length $\mathcal{O}(N^2)$, where $N = XY$ and X, Y denote the height and width of the BEV feature map. If we assume perception range $R = X/2 = Y/2$, 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 [86], which is of linear complexity with the input size $\mathcal{O}(2N + NK)$, where K is the total number of the sampled key ($K \ll N = XY$).



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

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

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

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

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 the size of BEV grids, it still can be a bottleneck when the perception range becomes large. Inspired by [59], we propose a method to further reduce the number of input queries from $N = XY$ to $N = N_k \ll XY$ by using features with top-k confidence. 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 $\mathrm{z}_q \in \mathbb{R}^{C\times XY}$ 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

For all tasks, we train the depth distribution network with a depth map obtained by projecting LiDAR points into the image view, following BEVDepth [32].

3D Detection and Tracking. For the 3D object detection task, we follow CenterPoint [80] to predict the center heatmap with anchor-free and multi-group head [85]. After, we perform 3D tracking by tracking-by-detection using velocity-based closest distance matching [80]. For training sparse aggregation setting, we filter LiDAR points outside of the 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.

BEV Segmentation. For the BEV segmentation task, we attach a convolutional decoder head to obtain the prediction map following CVT [82]. Given a BEV feature map from Multi-modal Feature Aggregation (MFA) layers, the segmentation head encodes to a latent representation and decodes back to the final output segmentation map, followed by a sigmoid layer. Our BEV segmentation network predicts a semantic occupancy grid of vehicles and drivable area, trained with a focal loss [37].

4. Experiments

4.1. Experimental Settings

Dataset and Metrics. We conduct experiments on nuScenes [2], which provides radar point clouds at scale. For 3D object detection and tracking, we use official metrics: mAP [11], NDS [2], and AMOTA [76], and we follow the settings proposed by LSS [52] for BEV segmentation. We refer the reader to nuScenes [2] and LSS [52] for details of metrics.

Implementation Details. For the camera stream, we adopt BEVDepth [32] 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 [23]. Unless otherwise specified, we follow standard practices [32] for implementation and training details. We accumulate previous three BEV feature maps with an interval of 1 second, similar to BEVFormer [33].

Our models are trained for 24 epochs with AdamW [44] optimizer in an end-to-end manner, unless otherwise specified. In addition to image and BEV data augmentation [32], we randomly drop sweeps and points for radar [27]. Inference time is measured on an Intel Core i9 CPU and RTX 3090 GPU with a single batch and FP16 precision. The full experimental settings are provided in Appendix.

Method	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS
CenterPoint-P†* [80]	L	Pillars	-	59.8	49.4	0.320	0.262	0.377	0.334	0.198	-
CenterPoint-V†* [80]	L	Voxel	-	65.3	56.9	0.285	0.253	0.323	0.272	0.186	-
BEVDet† [19]	C	R50	256 × 704	39.2	31.2	0.691	0.272	0.523	0.909	0.247	15.6
CenterFusion† [48]	C+R	DLA34	448 × 800	45.3	33.2	0.649	**0.263**	0.535	0.540	**0.142**	-
BEVDepth† [32]	C	R50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198	11.6
RCBEV4d† [83]	C+R	Swin-T	256 × 704	49.7	38.1	0.526	0.272	0.445	0.465	0.185	-
CRAFT† [24]	C+R	DLA34	448 × 800	51.7	41.1	0.494	0.276	0.454	0.486	0.176	4.1
SOLOFusion† [51]	C	R50	256 × 704	53.4	42.7	0.567	0.274	**0.411**	**0.252**	0.188	11.4
CRN	C+R	R18	256 × 704	54.3	44.8	0.518	0.283	0.552	0.279	0.180	**27.9**
CRN	C+R	R50	256 × 704	**56.0**	**49.0**	**0.487**	0.277	0.542	0.344	0.197	20.4
PETR† [40]	C	R101	900 × 1600	44.2	37.0	0.711	0.267	0.383	0.865	0.201	1.7
MVFusion† [78]	C+R	R101	900 × 1600	45.5	38.0	0.675	0.258	0.372	0.833	0.196	-
BEVFormer [33]	C	R101	900 × 1600	51.7	41.6	0.673	0.274	0.372	0.394	0.198	1.7
BEVDepth† [32]	C	R101	512 × 1408	53.5	41.2	0.565	0.266	**0.358**	0.331	0.190	5.0
SOLOFusion [51]	C	R101	512 × 1408	54.4	47.2	0.518	0.275	0.604	0.310	0.210	-
SOLOFusion† [51]	C	R101	512 × 1408	58.2	48.3	0.503	0.264	0.381	**0.246**	0.207	-
CRN	C+R	R101	512 × 1408	59.2	52.5	0.460	0.273	0.443	0.352	0.180	**7.2**
CRN‡	C+R	R101	512 × 1408	**60.7**	**54.5**	**0.445**	0.268	0.425	0.332	0.180	-

Table 1. **3D Object Detection** on nuScenes `val` set. 'L', 'C', and 'R' represent LiDAR, camera, and radar, respectively. *: results from MMDetection3D [4]. †: trained with CBGS. ‡: use test time augmentation.

Method	Input	Backbone	NDS↑	mAP↑	mATE↓
PointPillars [26]	L	Pillars	55.0	40.1	0.392
CenterPoint [80]	L	Voxel	67.3	60.3	0.262
KPConvPillars [71]	R	Pillars	13.9	4.9	0.823
CenterFusion [48]	C+R	DLA34	44.9	32.6	0.631
RCBEV [83]	C+R	Swin-T	48.6	40.6	0.484
PETR [40]	C	V2-99	50.4	44.1	0.593
MVFusion [78]	C+R	V2-99	51.7	45.3	0.569
CRAFT [24]	C+R	DLA34	52.3	41.1	0.467
BEVFormer [33]	C	V2-99	56.9	48.1	0.582
BEVDepth [32]	C	ConvNeXt-B	60.9	52.0	0.445
BEVStereo [30]	C	V2-99	61.0	52.5	0.431
SOLOFusion [51]	C	ConvNeXt-B	61.9	54.0	0.453
CRN	C+R	ConvNeXt-B	**62.4**	**57.5**	**0.416**

Table 2. **3D Object Detection** on nuScenes `test` set. V2-99 is pre-trained on external depth dataset DDAD [14].

4.2. Main Results

3D Object Detection. For a fair comparison with previous state-of-the-art 3D detection methods, we train our model only on 3D detection task and report `val` and `test` set results in Tables 1 and 2. Under various input image sizes and backbone settings, our CRN ranks 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 image input and backbone (256×704 and R18) already outperforms competitors with a large image input and backbone (BEVFormer [33] and BEVDepth [32] 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. CRN also outperforms the LiDAR method CenterPoint-P [80], demonstrating the potential of cost-effective camera and radar to replace LiDAR for autonomous driving. Qualitative results are provided in Fig. 5 and Appendix.

Method	Input	AMOTA↑	AMOTP↓	FP↓	FN↓	IDS↓
CenterPoint [80]	L	63.8	0.555	18612	22928	760
DEFT [3]	C	17.7	1.564	22163	60565	6901
QD-3DT [18]	C	21.7	1.550	**16495**	60156	6856
CC-3DT [12]	C	41.0	1.274	18114	42910	3334
Sparse4D [38]	C	51.9	1.078	19626	**32954**	1090
CRN	C+R	**56.9**	**0.809**	16822	41093	**946**

Table 3. **3D Object Tracking** on nuScenes `test` set.

Method	Input	Backbone	Image Size	Veh.	D.A.	FPS
LSS [52]	C	EffNetB0	128 × 352	32.1	72.0	25
FIERY [17]	C	EffNetB4	224 × 480	35.8	-	8
CVT [82]	C	EffNetB4	224 × 448	36.0	74.3	35
GKT [5]	C	EffNetB4	224 × 448	38.0	-	**45.6**
BEVFormer [33]	C	R101	900 × 1600	46.7	77.5	1.7
Simple-BEV [15]	C	R101	448 × 800	47.4	-	-
Simple-BEV [15]	C+R	R101	448 × 800	55.7	-	7.3
CRN	C+R	R50	256 × 704	**58.8**	**82.1**	24.8

Table 4. **BEV Segmentation** on nuScenes `val` set.

3D Object Tracking and BEV Segmentation. We further demonstrate the generalization performance of CRN on 3D object tracking and BEV segmentation tasks. As shown in Table 3, our tracking result outperforms all published camera-only methods on nuScenes `test` set. Also, ours not only significantly improves AMOTA but also reduces AMOTP and identity switches.

CRN consistently achieves state-of-the-art performance on BEV segmentation task as shown in Table 4. When compared to previous segmentation methods with a small image input and backbone [82, 5], ours performs significantly better while maintaining a real-time inference speed thanks to our semantically rich and spatially accurate BEV feature map from camera and radar. CRN also achieves higher performance at a much faster FPS than a large image input and backbone [33], demonstrating the effectiveness of fusion.

Input	RVT	*All* NDS	*All* mAP	*All* mATE	*Car* 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 5. Ablation of view transformation methods. LiDAR and radar are used only for transformation and not used for feature aggregation.

	Input	*All* NDS	*All* mAP	*All* mATE	*Car* mAP
CenterPoint	L	52.8	41.2	0.406	73.9
BEVDepth	C	43.9	33.2	0.716	50.4
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

Table 6. Ablation of feature aggregation methods. Note that MFA with RVT is our full model.

	Input	*Car* mAP [0,100)	[0,30)	[30,60)	[60,100)	FPS
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 7. Analysis over various perception ranges. Suffix -S denotes sparse aggregation, and we use 256×704 and R50 for all camera streams.

	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)

Table 8. Analysis of robustness using *Car* class mAP. Six view drops denote the single modality is entirely off.

4.3. Ablation Studies

We conduct ablation studies on `val` set with a 3D detection task. Unless otherwise specified, models use two frames of 256×704 image, R50 backbone, and are trained for 24 epochs without CBGS [85]. For thorough comparison, we additionally build three baseline detectors for camera – BEVDepth [32], point – CenterPoint [80], and camera-point – BEVFusion [42]. Details of baselines and additional ablation studies are provided in Appendix.

View Transformation. In Table 5, 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 6 shows the comparison between different feature aggregation methods. BEVFusion [42] 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.

4.4. Analysis

Scaling Up Perception Range. In Table 7, we extend the perception range from 51.2m to 102.4m and also increase the evaluation range twice correspondingly (see Appendix for details). Although CenterPoint [80] uses 10 LiDAR sweeps, points become extremely sparse as the range increases, and thus performance is significantly degraded at far distances. On 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 long-range perception. Moreover, CRN with sparse aggregation further improves the inference speed while preserving comparable performance.

Robustness. To systematically analyze the robustness of sensor failure cases, we randomly drop image and radar inputs in Table 8. 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 [6]. 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 unavailable, BEVFusion suffers from a performance drop over BEVDepth (-15.0%), while CRN still

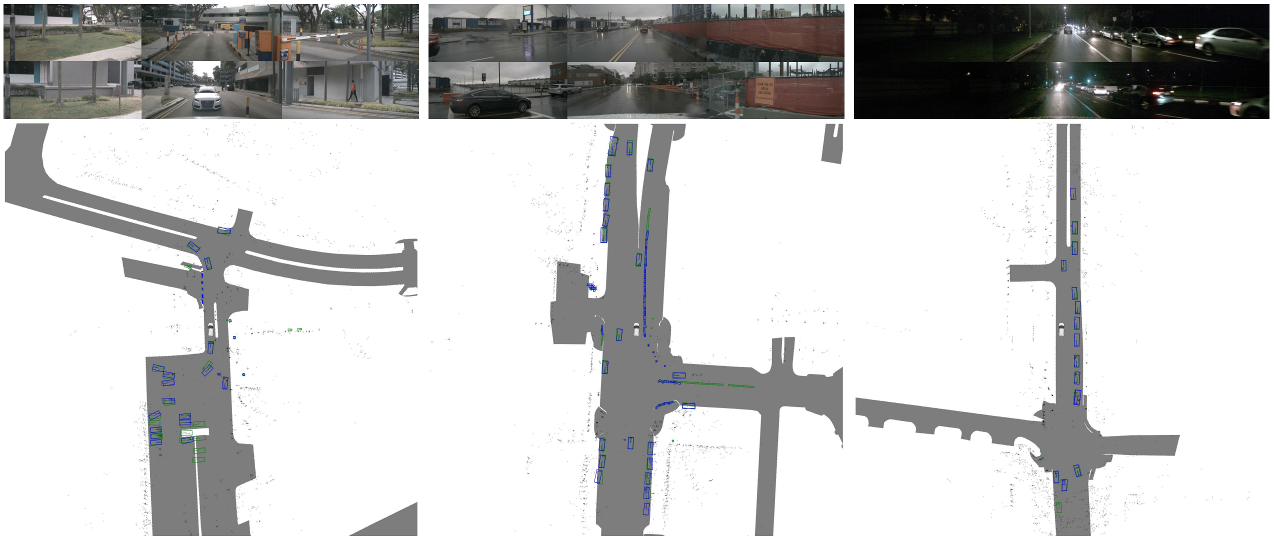


Figure 5. Qualitative results of long-range model on various weather conditions. Images on the top are the six camera views surrounding the vehicle. Green boxes are ground truths, blue boxes are our prediction results, and black dots are radar points. Perception ranges are set to $200m \times 200m$, and ground truth maps on the background are used for visualization. Best viewed in color with zoom in.

	Input	Sunny	Rainy	Day	Night
CenterPoint [80]	L	62.9	59.2	62.8	35.4
RCBEV [83]	C+R	36.1	38.5	37.1	15.5
BEVDepth [32]	C	39.0	39.0	39.3	16.8
CRN	C+R	**54.8**(+15.8)	**57.0**(+18.0)	**55.1**(+15.8)	**30.4**(+13.6)

Table 9. Analysis of different lighting and weather conditions using mAP metric. CenterPoint [80] results are from BEVFusion [42], and BEVDepth results are reproduced by us.

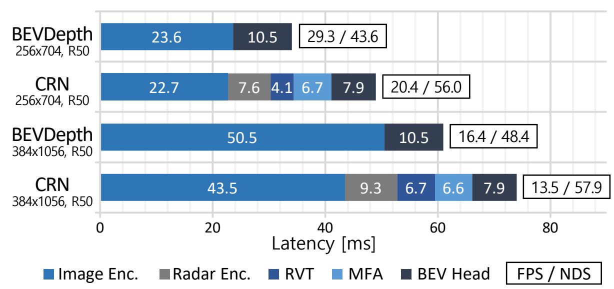


Figure 6. Inference time analysis of proposed components. All latency numbers are measured with batch size 1, GPU warm-up, and FP16 precision.

keeps the competitive performance (-5.6%). This advantage comes from our attention module, which can adaptively choose modalities to use.

Weather and Lighting. 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.

Inference Time. We analyze the inference time of each proposed component in Fig. 6. In 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.*, temporal stereo methods [30, 51]) 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.

5. Conclusion

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

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00236245, Development of Perception/Planning AI SW for Seamless Autonomous Driving in Adverse Weather/Unstructured Environment and 2022R1A2C200494412)

References

[1] 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)*, pages 1090–1099, 2022. 3

[2] 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. 1, 2, 3, 5

[3] Mohamed Chaabane, Peter Zhang, J Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 6

[4] 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. 6

[5] 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. 6

[6] 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. 3, 7

[7] 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)*, pages 1836–1843, 2014. 2

[8] Continental. Continental ARS 408-21 Datasheet. https://conti-engineering.com/components/ars-408/. Accessed: 2023-03-01. 1, 3

[9] Hao Dong, Xianjing Zhang, Xuan Jiang, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, Huimin Lu, Juho Kannala, and Xieyuanli Chen. Superfusion: Multilevel lidar-camera fusion for long-range hd map generation and prediction. *arXiv preprint arXiv:2211.15656*, 2022. 3

[10] Florian Drews, Di Feng, Florian Faion, Lars Rosenbaum, Michael Ulrich, and Claudius Gläser. Deepfusion: A robust and modular 3d object detector for lidars, cameras and radars. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 560–567, 2022. 3

[11] 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. 5

[12] Tobias Fischer, Yung-Hsu Yang, Suryansh Kumar, Min Sun, and Fisher Yu. Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 2294–2305, 2023. 6

[13] 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)*, pages 407–412, 2011. 2

[14] 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)*, pages 2485–2494, 2020. 2, 6

[15] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765, 2023. 6

[16] 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. 4

[17] 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. 1, 6

[18] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE Trans. Pattern Anal. Mach. Intell.)*, 45(2):1992–2008, 2022. 6

[19] 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. 1, 6

[20] 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. 2

[21] Don H Johnson and Dan E Dudgeon. *Array signal processing: concepts and techniques*. Simon & Schuster, Inc., 1992. 3

[22] 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)*, pages 388–402, 2020. 3

[23] 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)*, pages 10857–10864, 2020. 2, 3, 5

[24] 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. 2, 3, 6

[25] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5750–5757, 2018. 3

[26] 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, 4, 6

[27] Zhaoqi Leng, Guowang Li, Chenxi Liu, Ekin Dogus Cubuk, Pei Sun, Tong He, Dragomir Anguelov, and Mingxing Tan. Lidaraugment: Searching for scalable 3d lidar data augmentations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 7039–7045, 2023. 5

[28] Jian Li and Petre Stoica. *MIMO radar signal processing*. John Wiley & Sons, 2008. 3

[29] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4634, 2022. 2, 3

[30] 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. 4, 6, 8

[31] 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)*, 2022. 3, 4

[32] 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, 4, 5, 6, 7, 8

[33] 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. 2, 3, 5, 6

[34] 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. 3

[35] 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)*, pages 10233–10240, 2020. 3

[36] 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)*, pages 2117–2125, 2017. 4

[37] Tsung-yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 5

[38] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 6

[39] 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. 3

[40] 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)*, pages 531—548, 2022. 6

[41] 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. 4

[42] 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. 3, 4, 7, 8

[43] 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)*, pages 12507–12516, 2021. 3

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 5

[45] 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)*, pages 129–144, 2022. 2

[46] 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)*, pages 924–932, 2019. 3

[47] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *Proceedings of the European Radar Conference (EuRAD)*, pages 129–132, 2019. 2, 3

[48] 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. 2, 3, 6

[49] 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)*, pages 10386–10393. IEEE, 2020. 3

[50] 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)*, pages 3142–3152, 2021. 2

[51] 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. 2, 6, 8

[52] 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. 1, 2, 3, 4, 5, 6

[53] Alexander Popov, Patrik Gebhardt, Ke Chen, Ryan Oldja, Heeseok Lee, Shane Murray, Ruchi Bhargava, and Nikolai Smolyanskiy. Nvradarnet: Real-time radar obstacle and free space detection for autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6958–6964, 2023. 2

[54] 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)*, pages 918–927, 2018. 3

[55] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 2

[56] 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)*, pages 5105–5114, 2017. 2, 4

[57] 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)*, pages 8555–8564, 2021. 1, 2, 3, 4

[58] 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)*, pages 11138–11147, 2020. 2, 3

[59] 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. 5

[60] 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)*, pages 9200–9206, 2022. 2

[61] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10529–10538, 2020. 2

[62] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019. 2

[63] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1711–1719, 2020. 2

[64] Juyeb Shin, Francois Rameau, Hyeonjun Jeong, and Dongsuk Kum. InstaGraM: Instance-level Graph Modeling for Vectorized HD Map Learning. *arXiv preprint arXiv:2301.04470*, 2023. 1

[65] 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)*, pages 1991–1999, 2019. 2

[66] Liat Sless, Bat El Shlomo, Gilad Cohen, and Shaul Oron. Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019. 2

[67] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5725–5734, 2021. 2

[68] Peter Svenningsson, Francesco Fioranelli, and Alexander Yarovoy. Radar-pointgnn: Graph based object recognition for unstructured radar point-cloud data. In *Proceedings of the IEEE Radar Conference (RadarConf)*, pages 1–6, 2021. 2

[69] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J

Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 2

[70] 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)*, pages 9627–9636, 2019. 2

[71] Michael Ulrich, Sascha Braun, Daniel Köhler, Daniel Niederlöhner, Florian Faion, Claudius Gläser, and Holger Blume. Improved orientation estimation and detection with hybrid object detection networks for automotive radar. In *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 111–117, 2022. 2, 6

[72] 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. 4

[73] 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)*, pages 4604–4612, 2020. 3

[74] 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)*, pages 913–922, 2021. 2

[75] Yue Wang, Alireza Fathi, Abhijit Kundu, David A Ross, Caroline Pantofaru, Tom Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–34, 2020. 2

[76] Xinshuo Weng and Kris Kitani. A baseline for 3d multi-object tracking. In *arXiv preprint arXiv:1907.03961*, 2019. 5

[77] Rob Weston, Sarah Cen, Paul Newman, and Ingmar Posner. Probably unknown: Deep inverse sensor modelling radar. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 5446–5452, 2019. 2

[78] Zizhang Wu, Guilian Chen, Yuanzhu Gan, Lei Wang, and Jian Pu. Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 3, 6

[79] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337–3352, 2018. 4

[80] 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. 1, 3, 5, 6, 7, 8

[81] 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)*, pages 720–736, 2020. 3

[82] 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. 1, 2, 3, 5, 6

[83] 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. 3, 6, 8

[84] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2

[85] 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. 5, 7

[86] 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. 4