Learning 3D Occupancy from Beam Overlap in 2D Rotating mmWave Radar

Yu Du[1*], Ruifeng Nie[1*], Long Ma[1], Chengpei Xu[1,2], Yu Liu[1], Weimin Wang[1†]

[1]Dalian University of Technology, Dalian, China
[2]University of New South Wales, Sydney, Australia

Abstract

Robust 3D perception under adverse weather is critical for autonomous systems. While mmWave Radars are inherently weather-resistant, conventional 2D rotating Radar sensors lack direct elevation resolution, limiting their 3D perception ability. Although 4D imaging Radars can provide elevation information, they typically suffer from limited coverage and range. In this work, we exploit a key observation about mechanically rotating 2D mmWave Radars: in each sweep, an overlap exists between adjacent azimuth beam coverage due to the width of the main lobe, which makes the reflected intensity difference imply object materials and geometric shapes, including elevation. With this observation, we propose a method that learns 3D occupancy by disentangling bird's-eye view (BEV) layout and elevation estimation from one frame Radar scan. Specifically, we partition one sweep into two interleaved subsets, corresponding to overlapping beam directions, and utilize them to infer coarse geometric structure through spatial differences and intensity patterns. Extensive quantitative and qualitative evaluations on two real-world datasets demonstrate that our proposed method outperforms existing baselines. The codes will be publicly available.

Introduction

Occupancy-based 3D perception has shown effectiveness for modeling dense spatial structures, especially for outdoor perception (Shi et al. 2024). While LiDAR- and camera-based methods achieve strong performance for 3D occupancy prediction under good weather (Zhang, Zhu, and Du 2023; Huang et al. 2023; Wei et al. 2023; Pan et al. 2024; Tang et al. 2024; Wang et al. 2024; Hou et al. 2024), their performance usually degrades in adverse conditions, making Radar a more robust compensation. Thus, recent efforts (Ding et al. 2024; Ma et al. 2024) try to extend Radar perception to the occupancy generation domain. RadarOcc (Ding et al. 2024) leverages 4D Radar tensors for occupancy prediction but suffers from high computational cost and low field of view. LiCROcc (Ma et al. 2024) improves the situation via knowledge distillation from LiDAR-camera fusion networks but treats Radar returns as LiDAR-like points, which underutilizes Radar-specific properties such as reflectivity patterns and beam overlap.

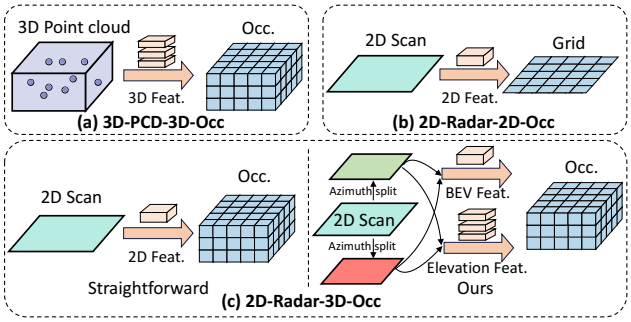


Figure 1: Illustration of different approaches for occupancy prediction. (a) 3D-PCD-based methods predict 3D occupancy from 3D point clouds with voxel-based backbones. (b) Conventional Radar-based methods extract 2D features to predict 2D occupancy. (c) Our method leverages structured 2D Radar scans by disentangling elevation cues from reflectivity intensity to predict 3D occupancy.

Among various Radar types used in autonomous driving, 2D mechanically rotating mmWave FMCW radars represent a widely adopted configuration for outdoor perception tasks, offering wide azimuthal coverage and long-range sensing. Different from sparse multi-channel Radars used in datasets like nuScenes (Caesar et al. 2020), or the limited field-of-view 4D Radars found in K-Radar (Paek, Kong, and Wijaya 2022), rotating 2D Radars provide full 360°scans with consistent angular sampling. These sensors have been applied to localization, odometry, and object detection (Hong, Petillot, and Wang 2020; Hong et al. 2022; Yin et al. 2022; Sakthi, Arvinte, and Vikalo 2023; Qian et al. 2021; Li et al. 2022a), owing to their robustness under adverse weather and their global scene coverage. However, their lack of vertical resolution limits their capacity for full 3D perception. Existing methods have attempted to overcome these issues via BEV feature learning, temporal fusion, and cross-modal enhancement (Qian et al. 2021; Li et al. 2022b, 2023; Wu et al. 2023), but most remain constrained to 2D representations.

Unlike LiDAR/4D-Radar-based 3D-3D (Cheng et al. 2021; Yan et al. 2021) or Radar-based 2D-2D (Weston et al.

*These authors contributed equally.
†Corresponding author: wangweimin@dlut.edu.cn.

2019; Kung, Wang, and Lin 2022) occupancy prediction tasks, this work aims to explore 3D prediction with 2D Radars, as illustrated in Figure 1. A straightforward approach is to extract 2D features and then use self-learning methods, such as MLPs or Transformers (Vaswani et al. 2017), to implicitly predict elevation. In this work, we revisit the 2D Radar sensing process and observe two underexploited sensing cues: (1) adjacent beams partially overlap due to the main lobe width exceeding the azimuth sampling interval, and adjacent beams exhibit spatial overlap; and (2) reflected intensity (RCS) varies with geometry and material. We leverage these cues to infer elevation from a single radar sweep, bridging the gap toward 2D Radar-based 3D occupancy prediction. The idea can be intuitively analogous to the depth estimation task with monocular and stereo cameras, where stereo recovers depth by leveraging disparities across overlapping observations of the same scenery.

To achieve 3D predictions, we design a BEV branch and an elevation branch to independently extract 2D semantic features (e.g., material and surface texture) and 3D spatial features (e.g., size and shape), as they relate to RCS. The BEV branch encodes two split streams separately and refines the extracted semantic features, providing detailed support for scene construction. BEV probabilities, guided by 2D annotations, are incorporated to reduce the impact of BEV factors on the elevation branch's spatial fitting. Specifically, we introduce an elevation transformer module that up-projects 2D BEV pillar into 3D voxel space to estimate elevation-wise occupancy distribution conditioned on 2D grid features. By disentangling BEV extraction and elevation estimation, our architecture can refine these aspects by aggregating them into a comprehensive 3D representation. The contributions of this work are:

- As far as we know, this is the first work to explore the task of 3D occupancy prediction from a single sweep of 2D rotating Radar, leveraging underutilized sensing properties for 3D scene perception.
- To achieve this, we propose an elevation-aware two-stream architecture that projects 2D Radar features into 3D space via structured beam split and transformer-based elevation inference, enabling effective 3D occupancy learning.
- We demonstrate the effectiveness of our method on two processed benchmarks based on real-world datasets, specifically adapted for the 2D Radar-based 3D occupancy task. Experimental results show consistent improvements over all baseline methods.

Related Work

LiDAR Occupancy Prediction

Several classic works focus on using LiDAR-only inputs and have shown strong performance. LMSCNet (Roldao, de Charette, and Verroust-Blondet 2020) employs a 2D U-Net (Ronneberger, Fischer, and Brox 2015) as its backbone by converting the elevation dimension into a feature channel, followed by a 3D segmentation head. This lightweight architecture strikes a balance between speed and accuracy. JS3C-Net (Yan et al. 2021) integrates the tasks of point cloud semantic segmentation and semantic scene completion. A point-voxel interaction module is designed to foster implicit mutual knowledge fusion between the two tasks. In contrast to voxel-based and point-based methods, PointOcc (Zuo et al. 2023) proposes a cylindrical coordinate system with a tri-perspective view (TPV) representation to process point cloud data. Inspired by TPVFormer (Huang et al. 2023), this approach projects each point onto three planes and then extracts features using a 2D backbone.

Radar Occupancy

Generation-based Radar Enhancement The inherent characteristics of Radar systems, including antenna size constraints and wave propagation properties, often result in lower-resolution data outputs and significant noise artifacts. To address these challenges, several methods (Yin et al. 2020; Zhang et al. 2024; Thilakanayake et al. 2024; Luan et al. 2024) have used high-precision LiDAR point clouds as supervisory signals to refine Radar data quality through generative frameworks. For instance, Thilakanayake et al. (Thilakanayake et al. 2024) propose a generative adversarial network (Goodfellow et al. 2020) method that improves Radar images using point cloud projection maps. RadarDiffusion (Luan et al. 2024) simulates the degradation of LiDAR point clouds into Radar images and learns the reverse process to recover high-resolution point clouds from degraded data. While these approaches push Radar data to mimic LiDAR's fine-grained details better, they also introduce an added complexity of learning the intrinsic characteristics of point clouds. In contrast, we utilize occupancy voxels to simplify sensors' observation representation, allowing Radar to gain deeper insights into the world's state instead of the principles of point clouds.

2D Grid Map. Occupancy grid (Moravec and Elfes 1985; Thrun et al. 2002) partitions the environment into grids with the occupancy state, assisting robotic navigation systems in obstacle recognition. Radar's long-range detection and all-weather operation capabilities make it particularly suited for this task. Traditional methods (Werber et al. 2015; Lombacher et al. 2017) rely on manually designed sensor characteristics but often struggle with complex Radar noise and varied environments. Recently, data-driven neural networks have emerged as a more effective alternative, as demonstrated in works like (Weston et al. 2019), (Kaul et al. 2020), and (Kung, Wang, and Lin 2022). Rob et al. (Weston et al. 2019) first introduce an inverse sensor model to convert raw Radar scans into 2D occupancy grids by LiDAR ground truth. Building on this foundation, Kung et al. (Kung, Wang, and Lin 2022) propose a polar sliding window approach to preserve the Radar's penetration capability.

3D Occupancy Prediction. A key challenge of extending a bird's-eye view (BEV) to 3D occupancy is that 2D Radar images lack elevation information. Weston et al. (Weston, Jones, and Posner 2021) attempt to predict elevation maps, but their approach highly relies on virtual data generated by the CARLA simulator (Dosovitskiy et al. 2017). In contrast, RadarOcc (Ding et al. 2024) directly utilizes dense

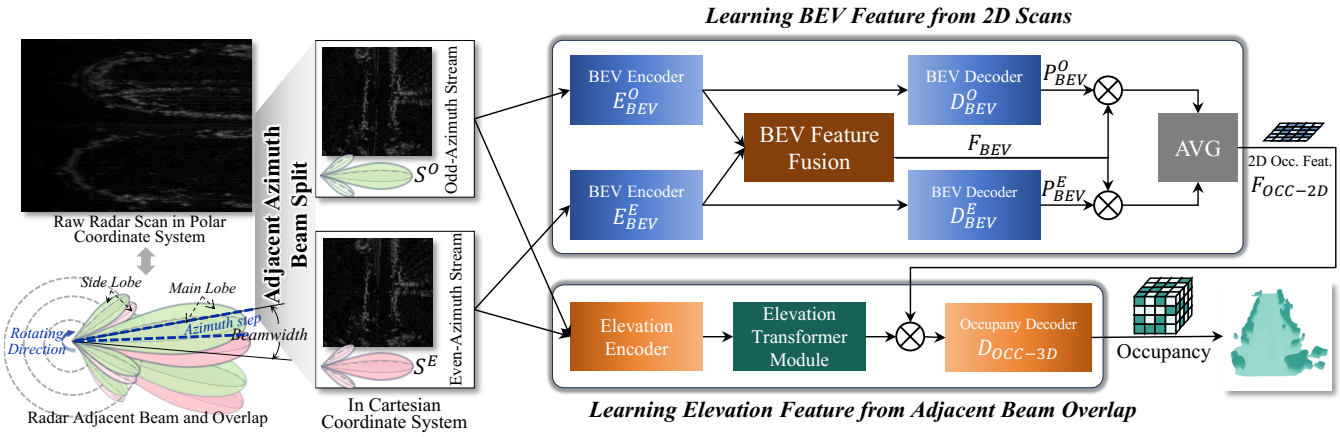


Figure 2: Overview of our proposed pipeline. Each Raw radar scan is first split into interleaved azimuthal streams: odd as S^O and even as S^E, which are separately processed by BEV encoders $E^{O/E}_{BEV}$ and jointly fed into the elevation encoder. The resulting BEV features of $E^{O/E}_{BEV}$ are fused into F_{BEV}, and corresponding 2D occupancy probabilities $P^{O/E}_{BEV}$ are obtained via decoders $D^{O/E}_{BEV}$. On the other hand, the elevation encoder and elevation transformer module produce an elevation probability map. The BEV features and occupancy predictions are combined as 2D occupancy features $F_{Occ\text{-}2D}$ and lifted into 3D voxel space by element-wise multiplication with the elevation probability map, followed by an occupancy decoder D_{OCC-3D} to predict the final 3D occupancy.

4D Radar tensor (4DRT) (Paek, Kong, and Wijaya 2022) and proposes a novel 3D occupancy prediction pipeline to capture low-reflective environmental details preserved in 4DRT. LiCROcc (Ma et al. 2024) addresses sparse Radar point clouds (Caesar et al. 2020) and introduces a knowledge distillation method that transfers information from LiDAR-camera fusion networks to Radar baseline models and Radar-camera fusion networks. Compared to redundant 4D tensors or overly filtered sparse points, scanning Radar specializes in 360° BEV imaging and longer distances, providing a wide perception range.

Motivated by these advantages, our work aims to recover elevation information from Radar Cross Section (RCS) values of individual lobes, establishing a novel paradigm for 3D occupancy prediction with 2D rotating Radar.

Method

Overview

To achieve the 3D occupancy prediction from the clues of the overlap of the adjacent beams and the intensity of RCS, we design a pipeline to disentangle 2D and elevation information, as shown in Figure 2. Based on the first observation, we split the raw scan into two interleaved azimuthal streams to isolate redundant spatial cues from overlapping beams, mitigating feature ambiguity and improving BEV representation. For the second observation, we disentangle elevation features from RCS data, and the adjacent sweep beams provide complementary dependencies for spatial modeling. For the second observation, we estimate elevation-related features by modeling RCS intensity variations across overlapping beams. The interleaved sweeps are jointly encoded to serve as semantic priors for implicit elevation features. These BEV and elevation features are then fused into 3D volumetric representations and decoded into voxel-wise occupancy predictions.

Disentanglement of BEV and Elevation

Main Lobe and Scanning Overlap. Radar emits directional millimeter-wave beams and captures their reflections from surrounding objects. These reflections are organized into a raw polar-coordinate scan image $I_\Pi \in \mathbb{R}^{A\times R}$, where Π indicates the polar coordinate system, the horizontal axis corresponds to Azimuth bins and the vertical axis to Range bins. Each pixel in I_Π represents the RCS intensity measured at a specific azimuth–range cell. Owing to the finite width of the Radar's main lobe, adjacent azimuth beams partially overlap, providing redundant spatial observations of the same object from slightly different viewpoints. For example, Navtech Radar's main lobe has a 1.8° field of view, and scans every 0.9°, causing overlap and potential implications for geometric shapes. To utilize this, we split the Radar data into two streams, S^O_Π and $S^E_\Pi \in \mathbb{R}^{A\times R}$, as shown in Figure 3, ensuring no beam overlap in each split stream. We reorganize data by:

$$\begin{aligned} S^O_\Pi(2a,r), \quad S^O_\Pi(2a{+}1,r) &= I_\Pi(2a,r), \\ S^E_\Pi(2a{+}1,r), \quad S^E_\Pi(2a{+}2,r) &= I_\Pi(2a{+}1,r), \\ a = 0,1,\ldots,A/2, \quad r &= 0,1,\ldots,R \end{aligned} \tag{1}$$

where r denotes the range index. Then, they are converted to Cartesian coordinate system with nearest neighbor interpolation, resulting in two input sweeps $S^O, S^E \in \mathbb{R}^{W\times L}$.

BEV Branch. The BEV encoders generate two feature maps of interleaved azimuth streams, which are structurally overlapped yet spatially complementary. To enhance reliable semantic cues and suppress noisy patterns, we apply

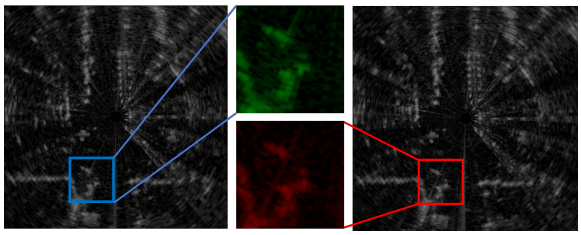
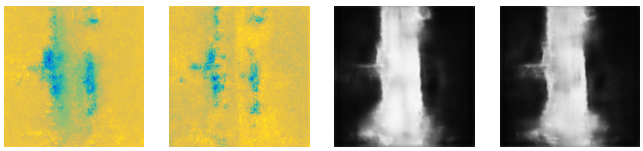


Figure 3: Difference between split adjacent odd and even streams in the Cartesian coordinate system. Due to the overlap coverage of Radar's main lobe, the two images are mostly identical in structure but differ in fine details.



Odd Feat. Even Feat. Guid. Even Guid. Odd

Figure 4: BEV features and guidance. Colors represent feature intensity and occupancy probability.

CBAM (Woo et al. 2018) to refine and fuse the two feature maps. The refined features are passed to two parallel 2D decoders to produce 2D occupancy maps for guidance, as shown in Figure 4, supervised by BEV ground truth projected from LiDAR. These predictions guide the elevation branch to focus on inferring elevation from RCS, without relearning BEV structure.

Elevation Transformer Module

Estimating elevation from mmWave reflections is challenging due to the weak correlation between RCS and vertical geometry. As shown in Figure 5, attention (Vaswani et al. 2017) is utilized to implicitly learn 3D relations, similar to CaDDN (Reading et al. 2021), which learns categorical distributions.

Row and Column Bars Strategy. Since treating each pixel as a token incurs a quadratic complexity ($O(n^2)$), we adopt a bar-token strategy that processes features along row and column streams separately, reducing the computational cost to $O(2n)$ while retaining spatial structure. Specifically, the row stream is formulated as follows:

$$\text{MultiHead}(q_w, f_w, f_w) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2}$$

$$\text{head}_i = \text{Attn}(q_w W_i^Q, f_w W_i^K, f_w W_i^V) \tag{3}$$

where q_w and f_w denote the w-th row query and features.

Column stream is processed similarly. Two probability maps are obtained via sigmoid and multiplied to form the final elevation map.

Polar Position Embedding. We propose a Radar-specific positional encoding using polar coordinates. Let $T(\cdot)$ be sine if i is odd, cosine if even, and we can define:

$$PE^R(x, y) = T\left(\frac{\sqrt{x^2 + y^2}}{\alpha^{2i/(0.5d)}}\right) \tag{4}$$

$$PE^A(x, y) = T\left(\frac{\arctan(y/x)}{\alpha^{2i/(0.5d)}}\right) \tag{5}$$

where α is a scale factor, d the dimension size. $PE^{2D} = [PE^A, PE^R]$ is added to the feature tokens. To encode elevation:

$$PE^Z(z) = T\left(\frac{\arctan(z)}{\alpha^{2i/d_{\text{elevation}}}}\right) \tag{6}$$

and the 3D embedding is injected into the elevation queries.

Feature Fusion for 3D Occupancy

We aggregate two BEV probability maps and their shared features using weighted averaging (dot product). The resulting 2D occupancy feature is multiplied with the elevation probability to produce the final 3D occupancy voxel tensor.

Losses

For 2D occupancy map regression, we apply Mean Squared Error (MSE) loss:

$$L_{\text{Occ-2D}} = \text{MSE}^O_{\text{BEV}} + \text{MSE}^E_{\text{BEV}} \tag{7}$$

3D occupancy loss includes Binary Cross-Entropy and Soft-IoU (Huang et al. 2020):

$$L_{\text{Occ-3D}} = L_{\text{BCE}} + L_{\text{Soft-IoU}} \tag{8}$$

The total training loss is:

$$L_{\text{total}} = L_{\text{Occ-2D}} + L_{\text{Occ-3D}} \tag{9}$$

Experiments

Setup

Raw Datasets. We evaluate our method on two real-world multimodal datasets with 2D scanning mmWave Radar and LiDAR: Oxford Radar RobotCar (**ORR**) with Navtech CTS350-X (Barnes et al. 2020) and **Boreas** with Navtech CIR304-H (Burnett et al. 2023), both operating at a 1.8° beam width and 0.9° azimuth resolution. For ORR, we follow the MVDNet (Qian et al. 2021) setup using 8,854 LiDAR-Radar pairs, split into 7,064 training and 1,790 testing frames with no geographical overlap. LiDAR voxels from 30 nearby frames are aggregated to generate dense occupancy supervision. The Boreas dataset includes the *Boreas-Objects-V1* sequence with 4,024 aligned Radar-LiDAR pairs and car annotations. We split the data into 70% training and 30% testing, and use 71-frame aggregated LiDAR to produce 3D occupancy ground truth.

Data Processing for Occupancy Labels. Similar to previous work (Tian et al. 2024; Li et al. 2024; Liu et al. 2024), we utilize a ray casting algorithm on the LiDAR point cloud to generate ground truth voxels, assigning each an occupancy state including the occupied, free and invalid. The invalid state is defined as the unknown one in invisible regions

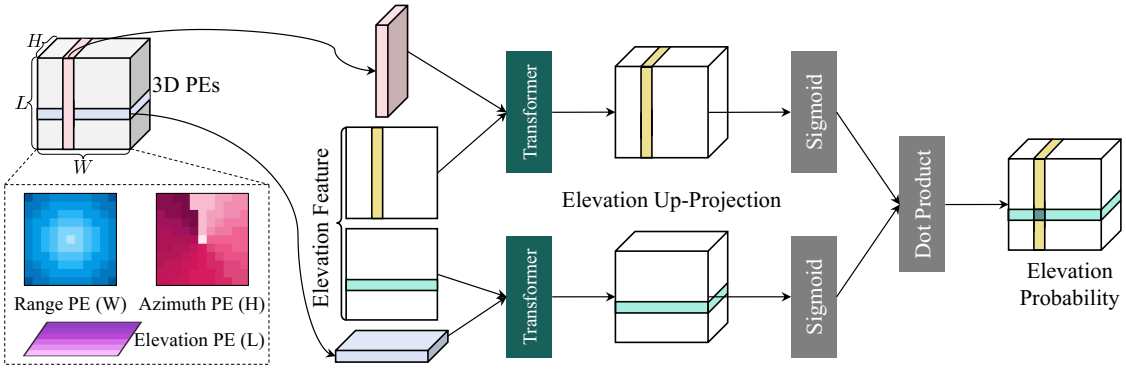


Figure 5: Architecture of the Elevation Transformer module, which takes the height feature map produced by the height encoder. Attention is used to extract elevation from RCS signals, which implicitly contain spatial structure. The Elevation feature part refers to the same height feature map, and is divided into row-wise and column-wise formats for two separate transformers.

where lasers are blocked by the obstructions ahead. These invalid voxels may be visible to Radar, which has a more powerful penetration ability. As such, we exclude them during both training and inference stages to prevent complex guidance for Radar prediction.

Then we aggregate multi-frame voxels into a multi-view occupancy map to further complete and refine the scene. Specifically, we merge temporally consecutive multiple LiDAR scans into a unified multi-view occupancy representation by taking the intersection of the voxel grids across frames. The overlay of multi-view observations from different origins expands the scope of the known region and even makes some obstructed objects perceivable, which helps provide more abundant supervision for Radar with long-range detection capability. In addition, we retain the dynamic vehicles inside the bounding boxes only in the reference frame of the aggregation sequence to prevent the dragging effects caused by their motion trails. Specifically, the region of interest is set to [-51.2m, 51.2m] $\times$ [-51.2m, 51.2m] $\times$ [-3.2m, 3.2m] for length, width and elevation, respectively. We crop and resize the Radar image within the ROI with a 512 x 512 resolution, and ground truth voxels of 512 x 512 x 32 with a 0.2m grid resolution.

Baseline Methods. To benchmark our method, we adapt and implement three representative architectures to the task of Radar-based 3D occupancy prediction: U-Net (Ronneberger, Fischer, and Brox 2015), PCF (Khurana et al. 2023), and Occ4Cast (Liu et al. 2024). Radar data are processed into 2D images or voxel grids, depending on the input requirements of each method.

U-Net, originally designed for high-resolution medical segmentation, is repurposed here by treating 3D occupancy estimation as a multi-label segmentation problem. To handle the possibility of multiple occupied elevations per pixel, we replace the standard cross-entropy loss with a binary loss across elevation bins. PCF reconstructs occupancy from spatio-temporal point cloud features, while Occ4Cast extends this with 3D convolutional architectures for occupancy completion and forecasting. We adopt its Conv3D variant for a fair comparison. We set the input/output temporal length to 1 to focus on scene completion. 2D Radar inputs are replicated along the vertical axis to generate voxelized inputs.

Metrics. We employ several widely adopted evaluation metrics to comprehensively evaluate model performance from different aspects. **Precision** and **Recall** measure the proportion of correctly predicted positive samples and the ability to recover all positive instances, respectively, while their harmonic mean is reported as the **F1 Score**. **Average Precision (AP)** evaluates performance across varying confidence thresholds, capturing the trade-off between precision and recall. The **Intersection over Union (IoU)** quantifies the spatial overlap between predicted and ground-truth occupancy regions, and serves as the primary evaluation metric in this work due to its strong correlation with occupyancy reconstruction quality.

Implementation Details. We use three ResNet50 (He et al. 2016) as encoders, simple 2D convolutions and 3D convolutions as 2D decoders and the 3D decoder. Following (Vaswani et al. 2017), the scale factor α for position embedding is set to 10000. The network is trained across 10 epochs on one V100 GPU. The training employs the Adam optimizer with the following settings: the batch size of 4, the initial learning rate set to $2e^{-5}$, and the decay factor of 0.2 every 4 epochs.

Quantitative Results

We use Boreas and ORR datasets to evaluate our method, as shown in Table 1. The quantitative results show that our proposed method outperforms other models, particularly in terms of IoU. Specifically, our method improves IoU by 3.14 and 2.92 on the Boreas and ORR datasets respectively, compared to the second U-Net method.

While some competing methods achieve higher precision, they significantly compromise on recall. This trade-off arises because LiDAR-based methods do not need to address significant missing elevation information, allowing them to focus primarily on reliable recognition. In contrast, our approach effectively mitigates the limitations of Radar data, leading to a better balance between precision and recall. As

Method \ Dataset	**Boreas** (Burnett et al. 2023)					**ORR** (Barnes et al. 2020)				
	Precision	Recall	F1 Score	AP	IoU	Precision	Recall	F1 Score	AP	IoU
U-Net (MICCAI'15)	24.53	29.40	26.27	19.56	15.19	39.99	42.40	40.71	34.80	25.68
PCF (CVPR'23)	**35.63**	14.75	20.24	22.73	11.30	**54.26**	26.17	34.98	39.39	21.28
Occ4cast (IROS'24)	32.59	15.77	20.53	21.47	11.49	53.15	30.98	38.69	**40.73**	24.09
Ours	23.24	**48.31**	**30.83**	**23.25**	**18.33**	42.16	**47.68**	**44.31**	40.64	**28.60**

Table 1: Performance comparison on the Boreas and ORR datasets. The best results are marked in bold.

a result, our approach shows improvements across multiple overall metrics, including the F1 score, AP and IoU.

Qualitative Results

We visualize our predicted results alongside those of baselines in Figure 6. Consistent with the quantitative findings, the qualitative results show that our method can predict a more complete scene, especially in smooth ground and adequate roadside structures. However, all methods struggle to generate fine details, especially on the roadside. A comparison between the Radar data and the predicted results reveals that the performance is inherently limited by the quality of the input data, as they tend to predict the occupancies that are captured in the original scans.

Ablation Study

In this section, we conduct a series of ablation experiments to demonstrate the effects of the designed components for the 3D occupancy learning with 2D Radar data.

Spatial Disentanglement. Our approach uses two branches to disentangle the BEV and elevation features. In order to analyze the effectiveness of this design, we replace the input with the raw Radar scans, using one encoder and two encoders, respectively. The results are shown in Table 2. Comparisons among the three setups demonstrate that multiple encoders working in concert outperform a single encoder learning all the knowledge. In addition, the comparison between raw Radar with two encoders and two sequential scanning sweeps with three encoders shows that observing the same scene from two perspectives through the scan segment strategy can provide rich RCS spatial modeling cues, ultimately enhancing perception performance.

Dataset	Input Types	#Enc	F1 ↑	AP ↑	IoU ↑
Boreas	Raw Image	1	30.36	**23.49**	17.98
	Raw Image	2	30.54	22.97	18.13
	Two sweeps	3	**30.83**	23.25	**18.33**
ORR	Raw Image	1	44.10	40.47	28.42
	Raw Image	2	43.85	40.27	28.26
	Two sweeps	3	**44.31**	**40.64**	**28.60**

Table 2: Ablation study on the input images w/ and w/o disentanglement on two datasets.

BEV Sup.	BEV Guid.	AP ↑	IoU ↑
✘	-	21.33	17.37
Raw Image	-	22.02	17.69
Occ-2D	✘	21.69	17.46
Occ-2D (Base)	✔	**23.25**	**18.33**

Table 3: Ablation study on BEV supervision and guidance. BEV supervision is employed to direct the decoded BEV occupancy probabilities, while BEV guidance refers to the fusion of the probability map with the BEV features.

BEV Supervision and Guidance. To make the elevation branch focus on parsing elevation information from Radar intensity, we introduce a BEV probability to obtain accurate 2D occupancy predictions in advance. The result of the impact by BEV supervision and guidance are shown in Table 3 This ablation experiment mainly consists of the following two adjustments compared to the base architecture: (1) remove 2D occupancy supervision for the BEV probability map or modify it to raw Radar scan; (2) cut off the path from BEV probability to BEV features which make the 2D probability no longer acts as a BEV guidance. The results show that adding BEV guidance supervised by BEV occupancy ground truth can bring 1.92 AP and 0.96 IoU improvement.

Types of Position Embedding. To investigate the effectiveness of polar PE, we conduct experiments using different PE formats by exploring four variants by combining the Polar PE and Cartesian PE with the inclusion or exclusion of z PE. As shown in Table 4, 3D polar PE improves the predictor accuracy, which can be attributed to the polar PE's ability to establish connections among data pixels that share similar azimuths or distances.

Application on Real-world Adverse Weather

To demonstrate the potential of mmWave Radar for 3D occupancy prediction under adverse weather, we additionally validate our model on the real-world RADIATE dataset (Sheeny et al. 2021), as shown in Figure 7. Due to the low quality of its LiDAR point clouds, which complicates the generation of dense voxel ground truth, we use the model trained on Boreas and perform inference directly on RADIATE. While cameras and LiDARs suffer from degradation and reduced visibility in challenging weather due to particles like water drops and snowflakes, Radar and predicted occupancy voxels maintain reliable performance.

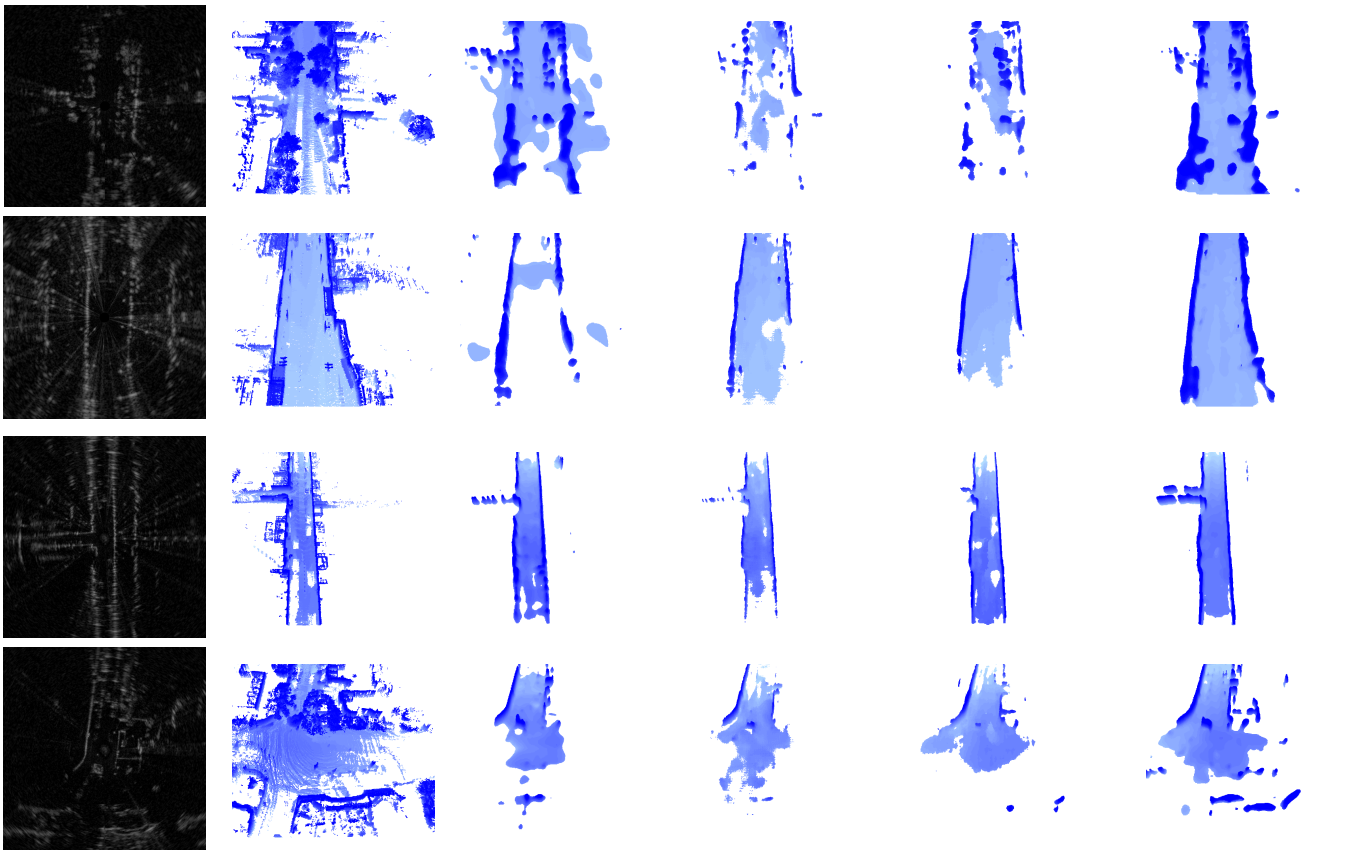


Figure 6: Visual comparisons of occupancy results on Boreas (Burnett et al. 2023) (top two rows) and ORR (Barnes et al. 2020) (bottom two rows). The ground truth (GT) occupancy labels are generated using LiDAR, which is elaborated in Sec. . The color of the voxels, ranging from light blue to dark blue, represents the elevation increasing from low to high. Compared to other baseline methods, our method predicts more detailed and accurate spatial structures.

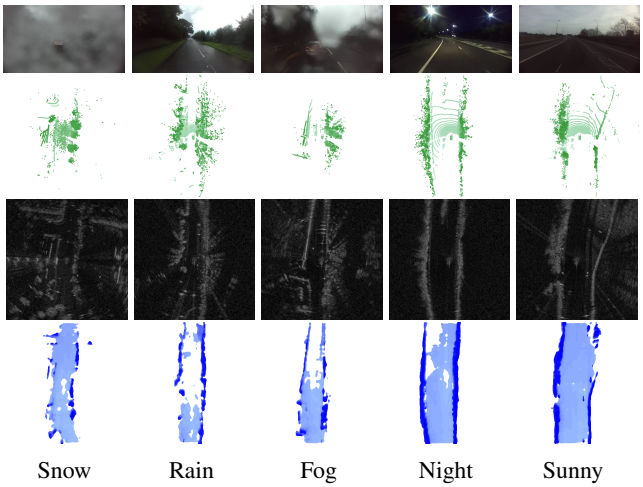


Figure 7: Qualitative results under adverse weathers on the RADIATE (Sheeny et al. 2021) dataset. From top to bottom: camera images, LiDAR points, Radar scans and our predictions. Camera and LiDAR data are for reference only.

PE	AP ↑	Δ	IoU ↑	Δ
Cartesian-2D	21.71	-1.54	17.56	-0.77
Cartesian-3D	22.08	-1.17	17.69	-0.64
Polar-2D	22.85	-0.40	18.23	-0.10
Polar-3D (Base)	**23.25**	-	**18.33**	-

Table 4: Ablation study on the type of position embedding.

Conclusion

We proposed a novel Radar-based 3D occupancy prediction approach that infers volumetric occupancy from 2D flat Radar scans. By leveraging two overlapping sequential Radar sweeps, our method disentangles spatial information from RCS signals. To reduce the complexity of elevation regression, BEV and elevation features are extracted independently and fused to generate final 3D predictions. We also constructed large-scale occupancy ground truth datasets and conducted extensive experiments. Results demonstrate that our method outperforms existing approaches and highlights the strong potential of Radar for robust 3D occupancy perception under various adverse weather conditions.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grant No. 62306059, 62506060 and 62472066.

References

Barnes, D.; Gadd, M.; Murcutt, P.; Newman, P.; and Posner, I. 2020. The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 6433–6438. IEEE.

Burnett, K.; Yoon, D. J.; Wu, Y.; Li, A. Z.; Zhang, H.; Lu, S.; Qian, J.; Tseng, W.-K.; Lambert, A.; Leung, K. Y.; et al. 2023. Boreas: A Multi-Season Autonomous Driving Dataset. *Int. J. Robot. Res.*, 42(1-2): 33–42.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 11618–11628.

Cheng, R.; Agia, C.; Ren, Y.; Li, X.; and Liu, B. 2021. S3cnet: A Sparse Semantic Scene Completion Network for LiDAR Point Clouds. In *Proc. Conf. Robot. Learn.*, 2148–2161. PMLR.

Ding, F.; Wen, X.; Zhu, Y.; Li, Y.; and Lu, C. X. 2024. Radarocc: Robust 3d occupancy prediction with 4d imaging radar. *Advances in Neural Information Processing Systems*, 37: 101589–101617.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative Adversarial Networks. *Commun. ACM*, 63(11): 139–144.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 770–778.

Hong, Z.; Petillot, Y.; Wallace, A.; and Wang, S. 2022. RadarSLAM: A robust simultaneous localization and mapping system for all weather conditions. *The international journal of robotics research*, 41(5): 519–542.

Hong, Z.; Petillot, Y.; and Wang, S. 2020. RadarSLAM: Radar Based Large-Scale SLAM in All Weathers. In *2020 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 5164–5170.

Hou, J.; Li, X.; Guan, W.; Zhang, G.; Feng, D.; Du, Y.; Xue, X.; and Pu, J. 2024. FastOcc: Accelerating 3D Occupancy Prediction by Fusing the 2D Bird's-Eye View and Perspective View. arXiv:2403.02710.

Huang, Y.; Tang, Z.; Chen, D.; Su, K.; and Chen, C. 2020. Batching Soft IoU for Training Semantic Segmentation Networks. *IEEE Sign. Process. Letters*, 27: 66–70.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9223–9232.

Kaul, P.; de Martini, D.; Gadd, M.; and Newman, P. 2020. RSS-Net: Weakly-Supervised Multi-Class Semantic Segmentation with FMCW Radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 431–436.

Khurana, T.; Hu, P.; Held, D.; and Ramanan, D. 2023. Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1116–1124.

Kung, P.-C.; Wang, C.-C.; and Lin, W.-C. 2022. Radar Occupancy Prediction With Lidar Supervision While Preserving Long-Range Sensing and Penetrating Capabilities. *IEEE Trans. Robot.*, 7(2): 2637–2643.

Li, P.; Wang, P.; Berntorp, K.; and Liu, H. 2022a. Exploiting Temporal Relations on Radar Perception for Autonomous Driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 17071–17080.

Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; et al. 2024. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 13333–13340. IEEE.

Li, Y.-J.; Hunt, S.; Park, J.; O'Toole, M.; and Kitani, K. 2023. Azimuth Super-Resolution for FMCW Radar in Autonomous Driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 17504–17513.

Li, Y.-J.; Park, J.; O'Toole, M.; and Kitani, K. 2022b. Modality-Agnostic Learning for Radar-LiDAR Fusion in Vehicle Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 918–927.

Liu, X.; Gong, M.; Fang, Q.; Xie, H.; Li, Y.; Zhao, H.; and Feng, C. 2024. Lidar-based 4d occupancy completion and forecasting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11102–11109. IEEE.

Lombacher, J.; Laudt, K.; Hahn, M.; Dickmann, J.; and Wöhler, C. 2017. Semantic Radar Grids. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 1170–1175.

Luan, K.; Shi, C.; Wang, N.; Cheng, Y.; Lu, H.; and Chen, X. 2024. Diffusion-Based Point Cloud Super-Resolution for mmWave Radar Data. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 11171–11177. IEEE.

Ma, Y.; Mei, J.; Yang, X.; Wen, L.; Xu, W.; Zhang, J.; Zuo, X.; Shi, B.; and Liu, Y. 2024. Licrocc: Teach radar for accurate semantic occupancy prediction using lidar and camera. *IEEE Robotics and Automation Letters*.

Moravec, H.; and Elfes, A. 1985. High Resolution Maps from Wide Angle Sonar. In *1985 IEEE Int. Conf. Robotics and Automation*, 116–121.

Paek, D.-H.; Kong, S.-H.; and Wijaya, K. T. 2022. K-radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions. *NeurIPS*, 35: 3819–3829.

Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; and Zhang, S. 2024. RenderOcc: Vision-Centric 3D Occupancy Prediction with 2D Rendering Supervision. In *2024 IEEE Int. Conf. Robotics and Automation (ICRA)*, 12404–12411.

Qian, K.; Zhu, S.; Zhang, X.; and Li, L. E. 2021. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary LiDAR and Radar Signals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 444–453.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8555–8564.

Roldao, L.; de Charette, R.; and Verroust-Blondet, A. 2020. Lmscnet: Lightweight Multiscale 3D Semantic Completion. In *2020 Int. Conf. 3D Vision (3DV)*, 111–119. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015, Part III*, 234–241. Springer.

Sakthi, M.; Arvinte, M.; and Vikalo, H. 2023. Automotive Radar Sub-Sampling via Object Detection Networks: Leveraging Prior Signal Information. *IEEE Open J. Intell. Transp. Syst.*, 4: 858–869.

Sheeny, M.; De Pellegrin, E.; Mukherjee, S.; Ahrabian, A.; Wang, S.; and Wallace, A. 2021. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 1–7. IEEE.

Shi, Y.; Jiang, K.; Li, J.; Qian, Z.; Wen, J.; Yang, M.; Wang, K.; and Yang, D. 2024. Grid-Centric Traffic Scenario Perception for Autonomous Driving: A Comprehensive Review. arXiv:2303.01212.

Tang, P.; Wang, Z.; Wang, G.; Zheng, J.; Ren, X.; Feng, B.; and Ma, C. 2024. SparseOcc: Rethinking Sparse Latent Representation for Vision-Based Semantic Occupancy Prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 15035–15044.

Thilakanayake, T.; De Silva, O.; Wanasinghe, T. R.; Mann, G. K.; and Jayasiri, A. 2024. A Generative Adversarial Network-based Method for LiDAR-Assisted Radar Image Enhancement. arXiv:2409.00196.

Thrun, S.; et al. 2002. Robotic mapping: A survey.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2024. Occ3d: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. *NeurIPS*, 36.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 6000–6010.

Wang, S.; Yu, J.; Li, W.; Liu, W.; Liu, X.; Chen, J.; and Zhu, J. 2024. Not All Voxels Are Equal: Hardness-Aware Semantic Scene Completion with Self-Distillation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 14792–14801.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. In *Int. Conf. Comput. Vis.*, 21729–21740.

Werber, K.; Rapp, M.; Klappstein, J.; Hahn, M.; Dickmann, J.; Dietmayer, K.; and Waldschmidt, C. 2015. Automotive Radar Gridmap Representations. In *2015 IEEE MTT-S Int. Conf. Microwaves for Intelligent Mobility (ICMIM)*, 1–4.

Weston, R.; Cen, S.; Newman, P.; and Posner, I. 2019. Probably unknown: Deep inverse sensor modelling radar. In *2019 international conference on robotics and automation (ICRA)*, 5446–5452. IEEE.

Weston, R.; Jones, O. P.; and Posner, I. 2021. There and Back Again: Learning to Simulate Radar Data for Real-World Applications. In *2021 IEEE Int. Conf. Robotics and Automation (ICRA)*, 12809–12816.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional Block Attention Module. In *Eur. Conf. Comput. Vis.*, 3–19.

Wu, Z.; Chen, G.; Gan, Y.; Wang, L.; and Pu, J. 2023. Mvfusion: Multi-View 3D Object Detection with Semantic-Aligned Radar and Camera Fusion. In *2023 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2766–2773. IEEE.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. In *AAAI*, volume 35, 3101–3109.

Yin, H.; Chen, R.; Wang, Y.; and Xiong, R. 2022. RaLL: End-to-End Radar Localization on Lidar Map Using Differentiable Measurement Model. *IEEE Trans. Intell. Transp. Syst.*, 23(7): 6737–6750.

Yin, H.; Wang, Y.; Tang, L.; and Xiong, R. 2020. Radar-on-Lidar: Metric Radar Localization on Prior Lidar Maps. In *2020 IEEE Int. Conf. Real-time Comput. Robot. (RCAR)*, 1–7.

Zhang, R.; Xue, D.; Wang, Y.; Geng, R.; and Gao, F. 2024. Towards dense and accurate radar perception via efficient cross-modal diffusion model. *IEEE Robotics and Automation Letters*.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-Path Transformer for Vision-Based 3D Semantic Occupancy Prediction. In *Int. Conf. Comput. Vis.*, 9433–9443.

Zuo, S.; Zheng, W.; Huang, Y.; Zhou, J.; and Lu, J. 2023. PointOcc: Cylindrical Tri-Perspective View for Point-based 3D Semantic Occupancy Prediction. *CoRR*.