

RadarGSOcc: Doppler-Aware 3D Semantic Occupancy Prediction Using 4D Radar Tensor with Geometric-Semantic Feature Extraction

Christofel Rio Goenawan, Dong-Hee Paek, *Member, IEEE*, and Seung-Hyun Kong, *Senior Member, IEEE*

Abstract—3D semantic occupancy prediction is crucial for autonomous driving, but camera-based methods degrade in low-light conditions and LiDAR-based methods are sensitive to adverse weather. 4D radar is more robust, yet its sparsity and the high cost of radar tensor processing limit its use. We propose Radar Geometric-Semantic Occupancy Prediction (RadarGSOcc), an efficient framework that directly processes 4D radar tensors by separately extracting geometric and semantic features, including Doppler information, with a deformable attention-based geometric refinement module as its key component. We also introduce the K-Radar 3DSemOcc benchmark, the first large-scale benchmark for fine-grained 3D semantic occupancy prediction using 4D radar tensors, built on the K-Radar dataset with refined LiDAR-based annotations. RadarGSOcc predicts 15 classes within a 51.2m range and achieves 38.2% IoU, 7.6% mIoU, and 25.5% weighted mIoU, a 16% relative improvement over RadarOcc in weighted mIoU (25.5 vs. 21.9; 23.8 vs. 20.9 averaged over three runs) with fewer parameters and FLOPs. The gain comes mainly from large static classes such as drivable area, vegetation, and terrain, whereas vehicles remain difficult: RadarOcc is better when car and truck/bus are merged into one class. RadarGSOcc also outperforms camera-based and camera-radar fusion baselines and narrows the gap with LiDAR-based methods. To assess robustness in adverse weather without per-frame ground truth, we further propose a proxy metric, Environment-Based KL Divergence (EBKL), under which 4D radar tensors yield more consistent predictions than camera and LiDAR in rain, sleet, and snow. The dataset and code will be publicly available.

Index Terms—3D semantic occupancy prediction, 4D radar tensor, adverse weather, autonomous driving, benchmark dataset, Doppler.

I. INTRODUCTION

3D semantic occupancy prediction enables autonomous vehicles to understand their 3D surroundings and identify safe navigable space for planning. By assigning an occupancy state and a semantic class to every voxel, it captures free space, drivable area, and objects of arbitrary shape that bounding-box detection cannot represent [1], [2]. While camera- and LiDAR-based methods perform well under ideal conditions, they inherit the limitations of their sensors: cameras lose most

of their information in low-light conditions, and LiDAR is severely degraded by rain, snow, and fog [3], [4].

4D radar provides a robust alternative. Millimeter-wave signals are largely insensitive to precipitation, fog, and illumination, and FMCW 4D radars measure the range, azimuth, elevation, and Doppler velocity of every reflection [5], [6]. The raw 4D radar tensor (4DRT) retains the full power spectrum before any thresholding [5], and its Doppler dimension directly encodes the radial velocity of each return, a strong cue for separating moving objects from static structures. Exploiting 4DRTs for dense prediction is nevertheless challenging: the tensor is large and noisy, so aggressive sparsification is required [7]; its low angular resolution blurs fine structures; and the Doppler dimension mixes power and velocity information that standard backbones do not readily consume [8].

Radar-based occupancy prediction has only recently emerged. RadarOcc [7] was the first to predict 3D occupancy directly from 4DRTs, and subsequent methods improve accuracy through LiDAR supervision [9] or camera-radar fusion [10], [11], but require additional sensors. More importantly, existing radar-only methods underutilize Doppler cues: when we reproduce RadarOcc on our benchmark, its performance remains nearly unchanged with or without the Doppler input (Table I).

In this paper, we propose **RadarGSOcc**, a Doppler-aware 3D semantic occupancy framework using 4D radar tensors. Our key idea is to extract the geometric (spatial power) and semantic (Doppler-related) information of the tensor separately and then fuse them. To this end, we introduce a **geometric-semantic feature extraction** module that processes the power and Doppler channels through separate sparse-convolution and deformable-attention branches [12], [13] before fusing them, and a lightweight **geometric refinement** module in the occupancy head that is designed to recover fine-grained structures, such as small objects, using the geometric features.

We also construct the *K-Radar 3DSemOcc* benchmark, the first large-scale benchmark for fine-grained 3D semantic occupancy prediction using 4D radar tensors, built from the K-Radar dataset [5]. It contains 14 semantic classes and an empty class over 5,910 training and 1,781 test frames. Labels are generated by a LiDAR-based occupancy model [1] and refined with the 3D bounding-box annotations of K-Radar, which more than doubles the number of moving-object voxels.

On this benchmark, RadarGSOcc achieves **7.6% mIoU** and **25.5% weighted mIoU** within a 51.2m range, improving the weighted mIoU of RadarOcc by **16% relative** (25.5 vs. 21.9) while using **11% fewer parameters** and **5% fewer FLOPs**; averaged over three training runs, the weighted mIoU is 23.8% vs. 20.9%. The gain comes mainly from large static classes

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Christofel Rio Goenawan is with the Robotics Program, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea (e-mail: christofel.goenawan@kaist.ac.kr).

Dong-Hee Paek was with the CCS Graduate School of Mobility, KAIST, Daejeon, Korea. He is now with the Department of Future Mobility, Korea University, Sejong, Korea (e-mail: dhpaek@korea.ac.kr).

Seung-Hyun Kong is with the CCS Graduate School of Mobility, KAIST, Daejeon, Korea (e-mail: skong@kaist.ac.kr). (*Corresponding author: Seung-Hyun Kong.*)

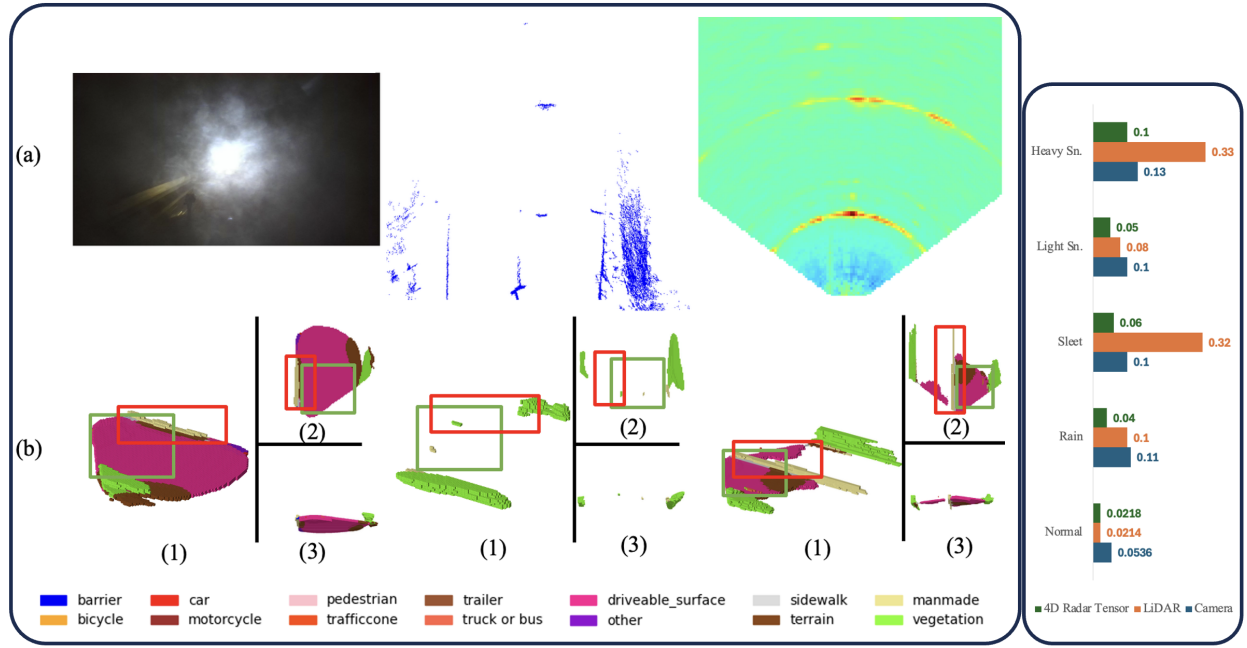


Fig. 1. (Left) Visualization of 3D semantic occupancy prediction under heavy snow in a highway scenario. (a) Input modalities: (left) front-view camera image, (middle) LiDAR point cloud, and (right) BEV of 4D radar tensor. (b) Predicted occupancy using (left) camera-based OpenOcc-C [1], (middle) LiDAR-based OpenOcc-L [1], and (right) the proposed RadarGSOcc with 4D radar input. RadarGSOcc produces more accurate predictions for key structures such as drivable areas (green box) and highway barriers (red box) under adverse weather. (Right) Quantitative comparison using the proposed EBKL proxy metric (lower is better). RadarGSOcc achieves lower EBKL, indicating more consistent predictions in adverse weather, while camera- and LiDAR-based methods degrade significantly.

such as drivable area, vegetation, and terrain. Vehicles remain difficult: RadarGSOcc improves truck/bus IoU but loses car IoU, and RadarOcc remains better when the two classes are merged into a single vehicle class. RadarGSOcc also outperforms camera-based [1], [14] and camera-radar fusion methods [1], [11], and narrows the gap with LiDAR-based approaches. Under adverse weather (e.g., rain, sleet, and snow), where camera and LiDAR degrade, its predictions remain more consistent. Since reliable ground truth is unavailable in such conditions, we propose a proxy metric, **Environment-Based KL Divergence (EBKL)**, which compares the predicted class distribution under a given weather condition with the normal-condition reference of the same environment. Under EBKL, 4D radar tensors give the lowest divergence of the three modalities in adverse weather.

In summary, the contributions of this paper are as follows:

- **K-Radar 3DSemOcc benchmark**: the first large-scale benchmark for fine-grained 3D semantic occupancy prediction using 4D radar tensors.
- **RadarGSOcc**: an efficient Doppler-aware framework whose geometric-semantic feature extraction and geometric refinement modules improve radar-based occupancy prediction with fewer parameters and FLOPs than RadarOcc.
- **Robustness and proxy metric**: a proxy metric (EBKL) for evaluating prediction consistency without ground truth, and an evaluation showing that 4D radar tensors give more consistent predictions than camera and LiDAR under adverse weather.

The remainder of this paper is organized as follows.

TABLE I
IMPACT OF DOPPLER INDEX ON RADAROcc PERFORMANCE (OUR REPRODUCTION ON THE K-RADAR 3DSEMOcc BENCHMARK).

Configuration	IoU	mIoU
RadarOcc (with Doppler)	36.4	6.7
RadarOcc (w/o Doppler)	35.9 (−0.5)	6.7 (0.0)

Section II reviews related work, Section III presents the benchmark and method, Section IV reports experiments, and Section V concludes the paper. Supplementary material on the benchmark, implementation details, and additional results is appended after the main text.

II. RELATED WORK

A. 4D Radar Datasets

Large-scale camera and LiDAR datasets have significantly advanced autonomous driving research [15], [16]. In contrast, 4D radar datasets remain limited despite their robustness in adverse weather [5], [17]–[19]. Early datasets such as Astyx [17] contain only 500 frames, while VoD [19] and TJ4DRadSet [18] mainly focus on urban scenes under clear conditions. K-Radar [5] addresses these limitations by providing 35,000 frames of 4D radar tensor data across diverse weather conditions, including rain, fog, and snow. However, it lacks annotations for 3D semantic occupancy prediction, motivating our proposed benchmark.

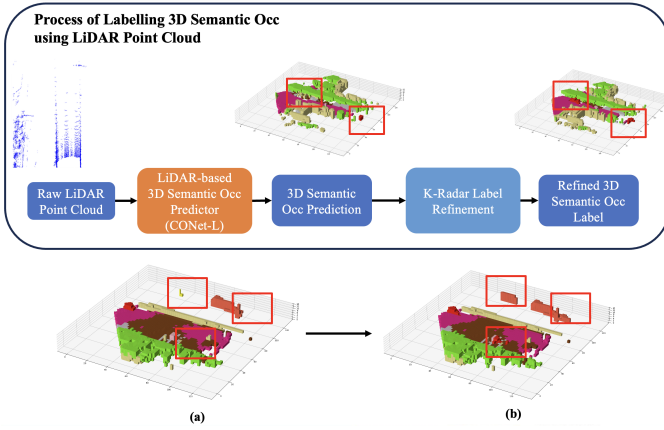


Fig. 2. Overview of the K-Radar 3DSemOcc labeling process. (a) Initial LiDAR-based labels. (b) Refined labels using K-Radar annotations, improving moving objects (red box).

B. 3D Semantic Occupancy Prediction

3D semantic occupancy prediction enables holistic scene understanding, including free space and drivable areas [20]. Occ3D [2] introduced the first large-scale benchmark on nuScenes [21]. Recent methods improve temporal modeling [22], point completion [23], multi-camera prediction [3], [24], and unified occupancy frameworks [1]. Other works focus on efficiency and robustness through Gaussian-based occupancy [25], [26], sparse prediction [27], [28], label-noise handling [29], and projection consistency [30].

C. 4D Radar-based Perception

Early 4D radar research mainly focused on object detection. Methods such as RPFA-Net [31] and SMURF [32] adapt LiDAR-based frameworks and reduce noise, while fusion-based approaches [18], [33], [34] improve performance but are typically evaluated in clear-weather conditions. For direct radar tensor processing, RTNH and RTNH+ [5], [35] learn radar-specific representations for 3D detection.

Recently, radar-based occupancy prediction has emerged. RadarOcc [7] first applies 4D radar tensors to 3D occupancy prediction using transformer-based architectures [2], [36]. Subsequent works improve performance through LiDAR supervision [9], camera-radar fusion [10], and multi-modal learning [11], [37], [38].

Despite these advances, existing methods often struggle with accurate semantic prediction and rely on computationally expensive transformer architectures. In contrast, we propose an efficient framework that leverages geometric-semantic feature extraction and lightweight refinement for 3D semantic occupancy prediction from 4D radar tensors.

III. METHOD

A. K-Radar 3DSemOcc Benchmark

There is no large-scale dataset with fine-grained 3D semantic occupancy labels for 4D radar tensors: RadarOcc [7] evaluates on K-Radar with occupancy labels that distinguish only foreground and background. To address this, we build

a new benchmark based on K-Radar [5]. Following prior work [1], [2], we generate pseudo-labels from LiDAR using a single-frame strategy due to inaccurate odometry [35].

The labeling pipeline has two stages:

(1) **Initial labeling:** The LiDAR-based CONet-L model of OpenOccupancy [1], trained on nuScenes [21], generates 17-class pseudo-labels. We select scenes 2, 5, 8, 9, 10, 14, 16, 17, and 20 for training, and 3, 6, and 15 for testing; all of these scenes are recorded under normal weather conditions, and their environment, time of day, and number of frames are listed in Table S-I of the supplementary material.

(2) **Label refinement:** The pseudo-labels are refined with the 3D bounding-box annotations of K-Radar by assigning the object class (e.g., car, truck/bus, pedestrian) to the voxels within each bounding box. This more than doubles the number of moving-object voxels in the training split, from about 365k to 768k (Tables S-II and S-III).

Fig. 2 shows an example in which refinement recovers cars and trucks. We further simplify labels by merging bus and truck classes and removing construction vehicles. Manual verification of 525 randomly sampled non-empty voxels (threshold 0.8) gives an accuracy of **91.8%** (482 correct); per-class results, including the moving-object classes, are reported in Table S-IV of the supplementary material together with the verification protocol. We retain voxels above a probability threshold, where 0.8 provides the best balance between quality and coverage (Section IV-D3).

The final *K-Radar 3DSemOcc* benchmark contains **5,910 training** and **1,781 testing frames** with 15 classes (14 semantic classes and the empty class), covering urban, university-campus, and highway environments under day and night conditions. The pipeline is general and can be extended to other datasets.

B. RadarGSOcc: Doppler-Aware 3D Semantic Occupancy Prediction Using 4D Radar Tensor

1) **4D Radar Tensor Preprocessing:** The overall architecture of RadarGSOcc is shown in Fig. 3. We follow the preprocessing strategy of RadarOcc [7]. Specifically, azimuth, range, and elevation are concatenated, and for each spatial element we extract the top-3 power values along the Doppler dimension together with their Doppler indices. To reduce computation, we keep only the top 250 values per range bin instead of using percentile-based sparsification [35]. This produces a sparse 4D radar tensor that preserves key motion and structural information.

The final representation has 8 channels: the top-3 power values, the top-3 Doppler indices, and the mean and variance of the power along the Doppler dimension (Section S-II of the supplementary material). This design follows RadarOcc while enabling efficient use of Doppler information for 3D semantic occupancy prediction.

2) **Geometric-Semantic Feature Extraction:** Extracting useful information from the Doppler dimension is challenging [8], since it contains both radar power and velocity-related information [5]. To address this, we design a feature extractor that separately learns geometric and semantic features from 4D radar tensors. Given the input tensor

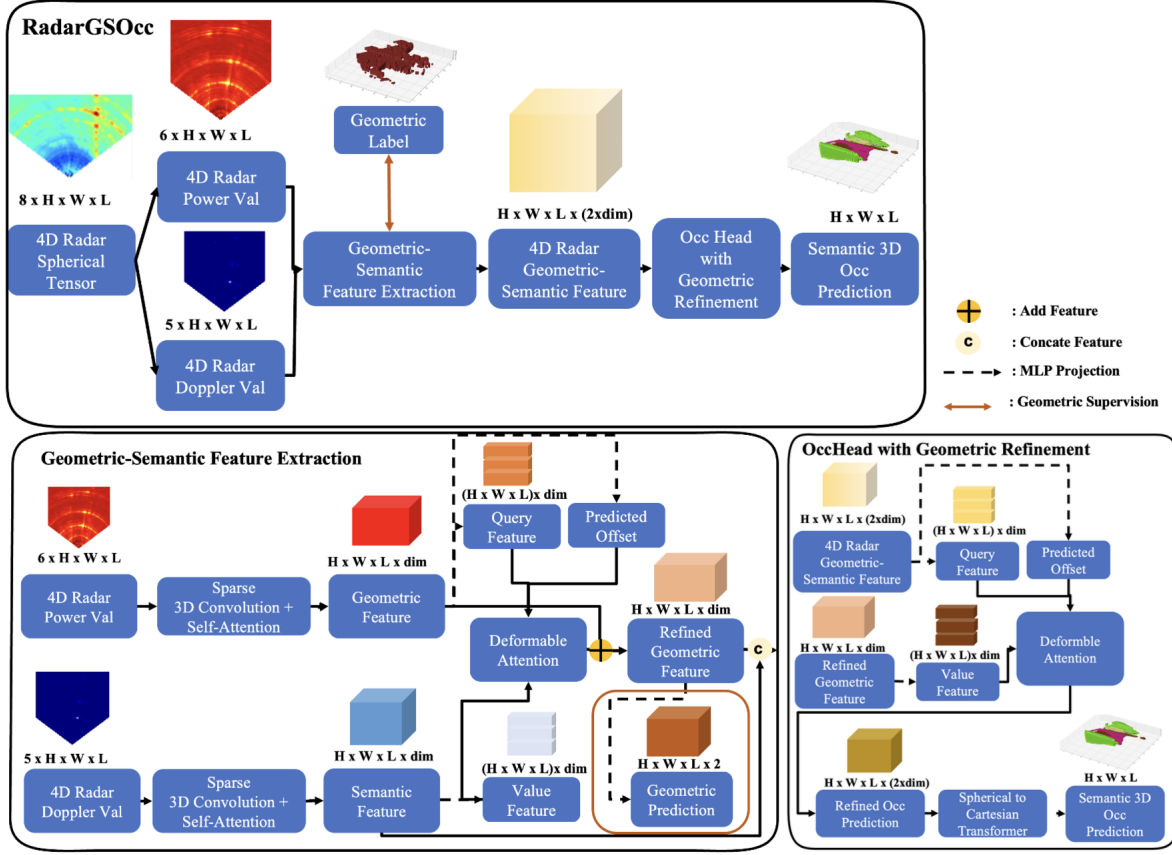


Fig. 3. Overview of the proposed RadarGSOcc architecture. The input 4D radar tensor is divided into power and Doppler components and processed by the proposed Geometric-Semantic Feature Extraction module (bottom left, Section III-B2), where $W \times H \times L$ denote spherical spatial dimensions and dim is the feature dimension. The extracted features are then passed to the occupancy head (bottom right), where the proposed geometric feature refinement module (Section III-B3) refines occupancy features using geometric representations. The final output is a 3D semantic occupancy prediction of size $H \times W \times L$.

$x_{input} \in \mathbb{R}^{8 \times H_{voxel} \times W_{voxel} \times L_{voxel}}$, we apply sparse 3D convolutions [12] to two channel groups:

$$x_{geo} = \text{Sparse3DConv}(x_{input}[6, :, :, :]), \quad (1)$$

$$x_{sem} = \text{Sparse3DConv}(x_{input}[3, :, :, :]). \quad (2)$$

Here, x_{geo} captures geometric features, while x_{sem} captures semantic and motion-related features. Channels 3–5 (the top-3 Doppler indices) are shared by both groups, since they are informative for both the geometric and the Doppler representations. Hence, the model can better exploit both power and Doppler information. Next, we enhance geometric features using a deformable self-attention mechanism inspired by [13]. This deformable attention focuses only on selected regions, making it more efficient than full attention methods [39], [40]. The 3D deformable self-attention yields the refined geometric feature x'_{geo} . The same process is applied to extract refined semantic features x'_{sem} . Implementation details of the deformable attention (number of sampling points and heads, offset normalization, and the interpolation scheme) are given in Section S-III of the supplementary material.

a) *Geometric-Semantic Fusion*: To further improve feature quality, we refine geometric features using semantic in-

formation via cross-attention, inspired by [8]. First, we project both features:

$$q_{fus} = \text{MLP}_q(x'_{geo}), \quad (3)$$

$$k_{fus} = \text{MLP}_k(x'_{sem}), \quad (4)$$

$$v_{fus} = \text{MLP}_v(x'_{sem}). \quad (5)$$

The normalized sampling offset Δ_{fus} is predicted as

$$\Delta_{fus} = \text{MLP}(x'_{geo}). \quad (6)$$

We then apply deformable attention:

$$A_{geo-sem} = \text{Softmax}\left(\frac{q_{fus} \cdot k_{fus}[\text{ref} + \Delta_{fus}]^T}{\sqrt{dim}}\right), \quad (7)$$

where ref denotes the normalized coordinates of the reference points. The refined geometric feature in spherical coordinates is then

$$x_{geo-refined} = A_{geo-sem} \cdot v_{fus} + x'_{geo}. \quad (8)$$

We also apply geometric supervision [3] by predicting whether each voxel contains an object:

$$x_{geo-supervised} = \text{MLP}(x_{geo-refined}). \quad (9)$$

Finally, we combine geometric and semantic features:

$$x_{geo-sem} = \text{concat}[x_{geo-refined}, x'_{sem}]. \quad (10)$$

The fused features are then processed using a feature pyramid network (FPN) head [41] and passed to the occupancy prediction head.

3) *OccHead with Geometric Feature Refinement*: The extracted geometric-semantic features are processed by a multi-scale 3D convolutional occupancy head, following [3]. However, similar to RadarOcc, the model tends to focus on large background regions (e.g., drivable area and terrain), while small objects (e.g., cars and trucks) are poorly captured due to loss of fine-grained features in earlier stages.

To address this issue, we introduce a **Geometric Feature Refinement** module based on 3D deformable attention [13], which is designed to enhance small-object representation using geometric features.

Let the occupancy feature be $x_{head} \in \mathbb{R}^{(H_{voxel} \times W_{voxel} \times L_{voxel}) \times dim}$ and the refined geometric feature be $x_{geo-refined}$, both in spherical coordinates. We project them into a shared latent space:

$$q_{ref} = MLP_{ref_q}(x_{head}), \quad (11)$$

$$v_{ref} = MLP_{ref_v}(x_{geo-refined}). \quad (12)$$

Deformable attention is then computed as:

$$A_{ref} = \text{Softmax} \left(\frac{q_{ref} \cdot v_{ref} [ref + \Delta_{ref}]^T}{\sqrt{dim}} \right). \quad (13)$$

The refined feature is obtained via residual fusion:

$$x_{head-refined} = MLP_{out}(A_{ref} \cdot v_{ref} + x_{head}). \quad (14)$$

Following [7], features are transformed from spherical to Cartesian coordinates using a 3D deformable transformer. Finally, two 3D convolution layers with ReLU activation [8] produce the output:

$$x_{pred} \in \mathbb{R}^{H_{voxel} \times W_{voxel} \times L_{voxel} \times N_{cls}}, \quad (15)$$

where N_{cls} is the number of semantic classes.

C. Loss Function

The network is trained in a fully supervised manner using ground-truth occupancy labels. Following [10], we adopt the cross-entropy loss L_{CE} as the primary objective. To address class imbalance, we additionally employ the Lovász-Softmax loss $L_{lovasz-softmax}$ [42]. To further improve geometric and semantic consistency, we incorporate geometric and class-wise affinity losses, L_{geo} and L_{sem} , as proposed in [43]. In addition, we apply a geometric supervision loss $L_{geo-feat}$ on the radar geometric features (see (9)) to enhance geometric feature learning; $L_{geo-feat}$ is a binary cross-entropy loss whose target is the binarized ground-truth occupancy (occupied or empty) of each voxel. All loss terms are weighted equally, and the overall training objective is

$$L_{overall} = L_{CE} + L_{lovasz-softmax} + L_{geo} + L_{sem} + L_{geo-feat}. \quad (16)$$

IV. EXPERIMENTS

A. Experiment Setup

We compare RadarGSOcc with the camera-, LiDAR-, radar point cloud-, and camera-radar fusion baselines of OpenOccupancy [1], with BEVFormer [14] and 4DRC-Occ [11], and with the 4D radar tensor-based RadarOcc [7]. For ablation, we evaluate RadarGSOcc variants without geometric refinement, without geometric-semantic fusion, and with standard 3D convolution refinement. For a fair comparison, RadarOcc and RadarGSOcc use the same feature dimension of 192. The prediction range is 0–51.2m forward, ± 25.6 m laterally, and -2.6 – 3.0 m vertically, with a voxel grid of $128 \times 128 \times 14$ and 0.4m resolution. Models are trained on the 5,910 frames of the training scenes and evaluated on the 1,781 frames of the test scenes, covering diverse urban and highway environments under day and night conditions. All 4D radar point cloud-, camera-, and LiDAR-based methods follow their original training settings. OpenOcc-L (64) and OpenOcc-L (128) denote the LiDAR-based model with 64- and 128-channel LiDAR inputs, respectively. Since the official code of 4DRC-Occ is not publicly available, we use our own reproduction of its version-A model, implemented in the OpenOccupancy framework with the hyperparameters reported in the original paper; because the original results are reported on a different dataset, the reproduction could not be checked against them.

B. Evaluation Metrics

We use binary IoU [7], mean IoU (mIoU) [44], and weighted mIoU (W-mIoU) [11], which weights each class by its voxel frequency so that extremely rare classes such as pedestrian and traffic cone (see Table S-III of the supplementary material) do not dominate the average:

$$\text{W-mIoU} = \sum_{c=1}^C \frac{n_c}{\sum_{i=1}^C n_i} \text{IoU}_c, \quad (17)$$

where IoU_c and n_c denote the IoU and the number of ground-truth voxels of class c , respectively, and $C = 14$ is the number of semantic classes.

Since accurate 3D semantic occupancy labels are difficult to obtain under adverse weather conditions due to degraded camera and LiDAR performance (Fig. 1), we introduce a supporting metric, **Environment-Based KL Divergence (EBKL)**:

$$x_{gt,env} = \text{Normalize} \left(\sum_{s=1}^{N_{scn}} x_{gt,s} \right), \quad (18)$$

$$\begin{aligned} \text{EBKL}(x_{pred,env,weather}) &= D_{KL}(x_{gt,env} \parallel x_{pred,env,weather}) \\ &= \sum_{c \in \mathcal{C}_E} x_{gt,env,c} \log \left(\frac{x_{gt,env,c}}{x_{pred,env,weather,c}} \right), \end{aligned} \quad (19)$$

where $x_{pred,env,weather} \in \mathbb{R}^{|\mathcal{C}_E|}$ denotes the normalized predicted class distribution under a specific environment and weather condition, and $x_{gt,env} \in \mathbb{R}^{|\mathcal{C}_E|}$ denotes the normalized *pseudo-label* class distribution aggregated from normal-condition scenes of the same environment. A small constant ϵ

is added to $x_{pred,env,weather,c}$ before normalization to avoid divergence when a class is absent from the prediction. In the K-Radar dataset, environments include urban, university-campus, and highway scenarios. EBKL assumes that the class distribution remains relatively consistent within the same environment, allowing distributional consistency under adverse weather to be evaluated using the divergence from the normal-condition distribution. Therefore, we compute EBKL only over a class subset $\mathcal{C}_E$ consisting of relatively stable background classes (drivable area, terrain, manmade, barrier, and vegetation) and frequently occurring moving-object classes (car and truck); both distributions are renormalized over $\mathcal{C}_E$. This assumption is examined in Section S-V of the supplementary material, which shows that highway scenes are well separated from the other environments, whereas the urban and university-campus distributions are close to each other. The assumption may not hold across datasets with substantially different scene distributions, such as K-Radar versus nuScenes [21] or Waymo [44]; therefore, EBKL may be less reliable for cross-dataset comparisons. Since EBKL measures only the divergence between voxel-class distributions and does not evaluate voxel-level semantic accuracy, it cannot replace IoU-based metrics: in particular, a degenerate predictor that outputs the normal-condition class distribution regardless of its input would obtain $EBKL = 0$. We therefore use EBKL only as a *proxy metric* for prediction consistency when ground-truth occupancy labels are unavailable under adverse weather, always together with the non-empty voxel ratio (Table VI) and qualitative results (Fig. 1); its use is supported by its strong negative correlation with mIoU across models under normal conditions (Fig. 6).

C. Implementation Details

All 4D radar-based models are trained for 15 epochs on 4 RTX 3090 GPUs using PyTorch 2.0.0 and Ubuntu 18.04. We use a batch size of 1 and a learning rate of 3×10^{-4} . The training objective follows (16). Camera-based [1], [14], fusion-based [1], [11], and LiDAR-based methods are trained with their original training settings.

D. Experiment Results

1) *Result of Proposed RadarGSOcc against State-of-the-Art Methods*: As shown in Table II, RadarGSOcc reaches 25.5% W-mIoU, a 16% relative improvement over RadarOcc [7] (21.9%), with about 11% fewer parameters and 5% fewer FLOPs (Table IV), although its frame rate is slightly lower (3.1 vs. 3.3 FPS). The improvement comes mainly from the large static classes that dominate the voxel-weighted average: drivable area (42.2% vs. 40.4%), vegetation (21.2% vs. 16.3%), and terrain (23.9% vs. 18.0%). For vehicles, the result is mixed. RadarGSOcc attains the highest truck/bus IoU of all methods (4.7%), but its car IoU drops from 6.3% to 0.6%, and when car and truck/bus are merged into a single vehicle class, RadarOcc remains better than RadarGSOcc (6.5% vs. 2.6% IoU). The higher truck/bus IoU therefore does not translate into better vehicle occupancy overall, and vehicle prediction from 4D radar tensors remains an open problem. Removing

the channel overlap reduces W-mIoU by 3.7 points (25.5% $\rightarrow$ 21.8%) and truck/bus IoU from 4.7% to 2.1%, while recovering some rare small-object classes (e.g., trailer and pedestrian; Table S-VI), indicating that the shared Doppler-index channels mainly benefit the dominant classes.

RadarGSOcc also outperforms the camera-based and camera-radar fusion baselines [1], [11], [14] in IoU, mIoU, and W-mIoU. Qualitative examples are shown in Fig. 4.

All methods still struggle on small foreground classes such as cones, pedestrians, and motorcycles because these classes each account for less than 0.1% of voxels (Table S-III), leading to severe class imbalance.

2) *Training with Different Random Seeds*: We train RadarGSOcc and the baseline RadarOcc [7] three times with different random seeds to evaluate training robustness (Table S-VIII of the supplementary material). Averaged over the three runs, RadarGSOcc achieves $37.43\% \pm 2.34\%$ IoU, $7.14\% \pm 0.41\%$ mIoU, and $23.83\% \pm 2.15\%$ W-mIoU, compared with $36.26\% \pm 0.23\%$, $6.72\% \pm 0.28\%$, and $20.93\% \pm 0.90\%$ for RadarOcc. RadarGSOcc is therefore better on average, but not in every run: in the third run, its IoU and mIoU (34.8% and 6.8%) are lower than those of RadarOcc (36.0% and 7.0%), although its W-mIoU remains higher (21.4% vs. 20.8%). RadarGSOcc also shows a much larger run-to-run variation (e.g., ± 2.15 vs. ± 0.90 in W-mIoU), so its training is less stable than that of RadarOcc. Table II and the other tables report the first run of each model, which is also the best of the three runs in W-mIoU for both models.

3) *Effect of Class Probability Threshold*: We evaluate different label-confidence thresholds (Table III). Since the threshold defines the evaluation label set itself, scores across rows are not directly comparable and only characterize the quality-coverage trade-off: lower thresholds keep noisier pseudo-labels, higher ones discard valid voxels. At 0.8, the manual verification gives a label accuracy of 91.8% (Section III-A); we adopt this threshold for all main experiments.

4) *Result of Proposed Geometric-Semantic Feature Extractor and Geometric Refinement*: We conduct ablation studies to evaluate geometric-semantic fusion and geometric refinement (Table IV), in which RadarOcc processes all input channels jointly and applies no refinement in the dense head. Separating the geometric and semantic branches alone (first row) already matches RadarOcc (6.9% vs. 6.7% mIoU) with fewer parameters and FLOPs. Neither module helps on its own: the deformable geometric refinement without fusion changes W-mIoU only marginally (21.8% $\rightarrow$ 21.1%), and geometric-semantic fusion without refinement slightly degrades it (20.7%), while refining the fused features with standard 3D convolution degrades it further (18.1%), indicating that a local convolutional refinement cannot exploit the fused geometric features. Only the combination of fusion and deformable-attention refinement yields a clear gain, raising W-mIoU to 25.5% (+3.7 over the base model and +7.4 over convolutional refinement; 23.8% on average over three runs, Section IV-D2) at a cost of only +0.4M parameters and +13 GFLOPs. The two modules therefore appear complementary: fusion injects Doppler-informed cues into the geometric features, and the deformable refinement allows the occupancy

TABLE II

3D SEMANTIC OCCUPANCY RESULTS ON THE K-RADAR 3DSEMOCC BENCHMARK (TEST SPLIT). BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE SHOWN IN BOLD, UNDERLINED, AND ITALIC, RESPECTIVELY. EBKL IS COMPUTED UNDER NORMAL WEATHER CONDITIONS (LOWER IS BETTER); “–” DENOTES NOT EVALUATED. “+ w/o overlap” DENOTES RADARGSOCC WITHOUT THE DOPPLER CHANNELS SHARED BETWEEN THE GEOMETRIC AND SEMANTIC BRANCHES. RADAROcc AND RADARGSOCC ROWS ARE THE FIRST OF THREE TRAINING RUNS (SECTION IV-D2). RESULTS FOR ALL CLASSES ARE GIVEN IN TABLE S-VI OF THE SUPPLEMENTARY MATERIAL.

Method	Sensor	IoU	mIoU	W-mIoU	EBKL	DA	Veg	Man	Car	Truck	Ter	Side
CONet-C [1]	Cam	16.9	3.2	11.3	0.093	27.4	4.2	0.0	0.6	0.0	11.4	0.0
BEVFormer [14]	Cam	27.3	4.1	16.0	0.072	34.0	10.4	0.1	0.0	0.0	11.9	0.5
OpenOcc-C [1]	Cam	30.0	5.4	17.1	0.054	34.7	17.3	<i>13.7</i>	<u>5.9</u>	1.8	19.0	0.9
OpenOcc-R [1]	Radar	26.0	4.0	14.7	0.054	34.0	8.8	5.1	1.2	0.0	8.9	0.8
OpenOcc-RC [1]	R+Cam	25.2	3.8	14.0	0.075	34.9	5.6	0.8	0.0	0.1	11.0	1.4
4DRC-Occ [11]	R+Cam	30.2	4.6	17.6	0.054	35.2	12.8	0.1	0.2	0.0	15.3	1.9
OpenOcc-L (64) [1]	LiDAR	<u>53.1</u>	<u>10.9</u>	<u>36.1</u>	0.021	<u>50.5</u>	<u>39.9</u>	<u>15.4</u>	<u>3.3</u>	<u>2.4</u>	<u>26.1</u>	9.1
OpenOcc-L (128) [1]	LiDAR	54.3	11.1	38.9	–	51.5	43.0	15.4	1.5	3.3	30.3	6.4
RadarOcc [7]	RadarT	36.4	6.7	21.9	0.032	40.4	16.3	10.3	6.3	0.6	18.0	2.4
RadarGSOcc(Ours)	RadarT	38.2	7.6	25.5	<u>0.022</u>	<u>42.2</u>	<u>21.2</u>	11.6	0.6	4.7	<u>23.9</u>	2.5
+ w/o overlap	RadarT	35.7	6.8	21.8	–	34.5	17.5	11.9	0.6	2.1	21.2	4.9

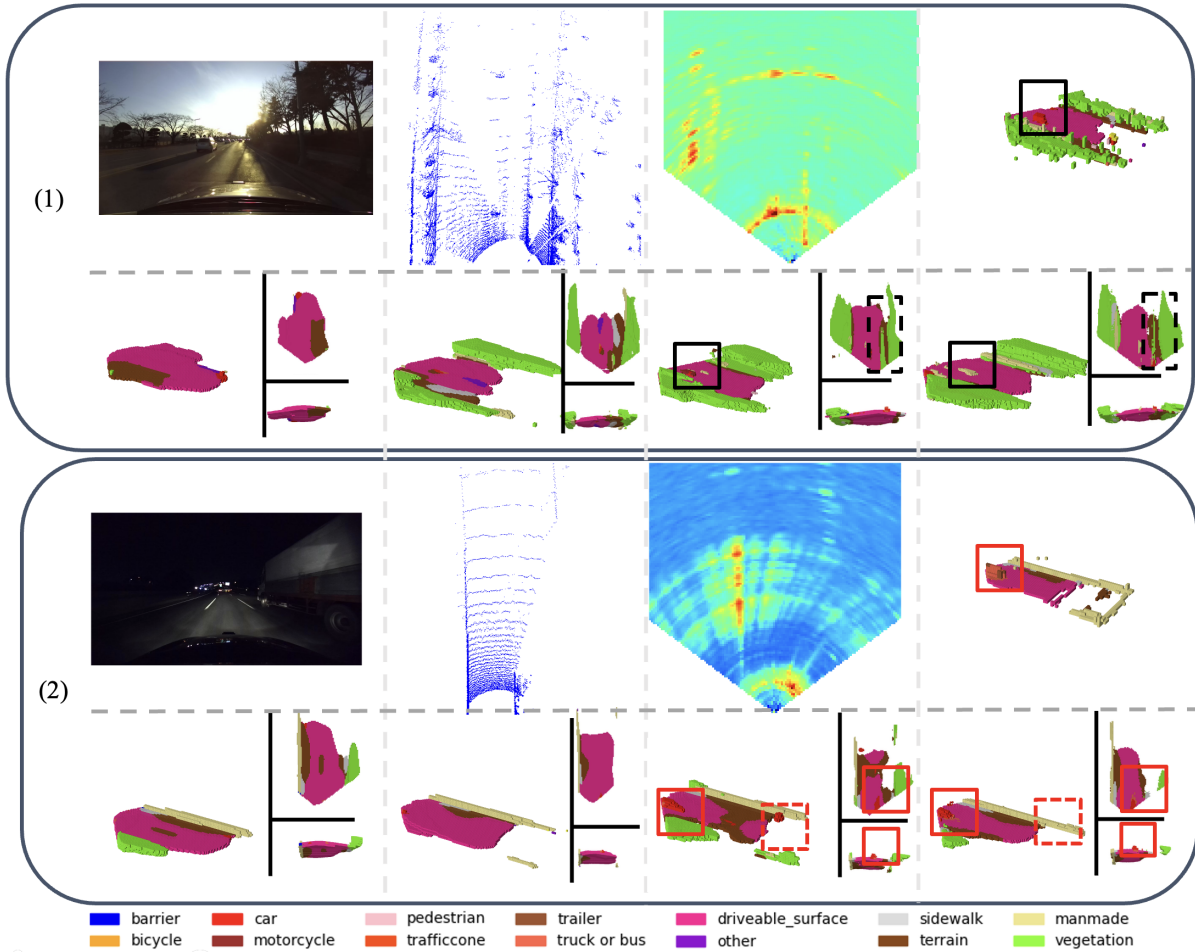


Fig. 4. Visualization of 3D semantic occupancy prediction using LiDAR (OpenOcc-L) [1], camera (OpenOcc-C) [1], and 4D radar tensors with RadarOcc [7] and the proposed RadarGSOcc in urban (1) and highway (2) scenarios. In each scenario, the top row shows (from left) the front-camera image, LiDAR point cloud, 4D radar tensor, and ground truth, while the bottom row shows camera-based prediction, LiDAR-based prediction, RadarOcc, and RadarGSOcc. Each prediction includes a 3D front-right view, BEV view, and front view. In these examples, RadarGSOcc predicts the highlighted moving objects (red and black boxes) more accurately, improves terrain and drivable area prediction (dashed black box), and reduces false moving-object predictions (dashed red box); quantitatively, however, its merged vehicle IoU is lower than that of RadarOcc (Section IV-D1).

head to exploit them. Because each variant is trained once and the run-to-run standard deviation of RadarGSOcc is about 2 points in W-mIoU (Section IV-D2), differences of about one point among the other variants should not be over-interpreted.

The geometric refinement module slightly lowers the frame rate (3.2 $\rightarrow$ 3.1 FPS) but improves W-mIoU. In the examples of Fig. 5, the refined features respond more strongly to moving objects (cars and trucks) and road structures.

TABLE III
PERFORMANCE OF RADARGSOCC ON THE K-RADAR 3DSEMOCB BENCHMARK UNDER DIFFERENT CLASS PROBABILITY THRESHOLDS.

Threshold	mIoU	W-mIoU	#Voxel GT	#Voxel Prediction
0.2	3.97	12.7	26,647	14,842
0.4	5.44	17.7	25,796	17,130
0.8	7.60	25.5	14,606	9,930
0.9	7.14	24.5	10,559	7,049

TABLE IV
ABLATION STUDY OF RADARGSOCC. FUSION: GEOMETRIC-SEMANTIC FUSION; CONV/DEF: GEOMETRIC REFINEMENT WITH 3D CONVOLUTION / DEFORMABLE ATTENTION. ✓ INDICATES ENABLED. PARAM IN MILLIONS (M) AND FLOPS IN GFLOPS (G). EACH VARIANT IS A SINGLE TRAINING RUN.

Fusion	Conv	Def	IoU	mIoU	W-mIoU	Param	FLOPs	FPS
RadarOcc [7]			36.4	6.7	21.9	117.5	1334.1	3.3
×	×	×	35.7	6.9	21.8	104.2	1260.1	3.2
×	×	✓	36.1	7.0	21.1	104.6	1270.8	3.1
✓	×	×	35.9	6.6	20.7	104.2	1262.7	3.1
✓	✓	×	32.3	5.6	18.1	105.0	1285.4	3.2
✓	×	✓	38.2	7.6	25.5	104.6	1273.5	3.1

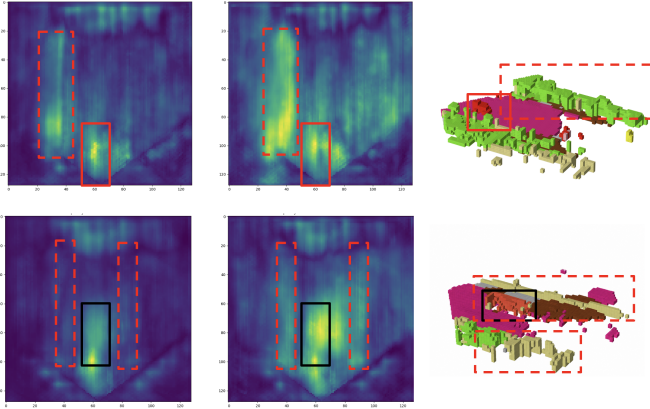


Fig. 5. Visualization of 3D occupancy features before (left) and after (middle) geometric refinement, compared with ground truth (right). In these examples, the refined features respond more strongly to moving objects (cars and trucks in red and black boxes) and road structures (dashed red boxes).

5) *Performance with Various Radar Tensor Processing and Top Doppler Dimension:* We evaluate different Top-K values and Doppler dimensions (Table V). The best configuration uses Top-K = 250 with all three Doppler components, reaching 25.5% W-mIoU. The effect of the Doppler dimension is not strictly monotonic at either Top-K, but the third Doppler component is clearly beneficial only at the higher Top-K = 250 (20.2% → 25.5%), where denser measurements make the additional component informative; at Top-K = 150 it instead adds noise (21.2% → 20.4%). This indicates that the benefit of finer Doppler resolution depends on sufficient measurement density. Across these Top-K and Doppler settings, FLOPs and frame rate remain nearly unchanged (Table V), as they affect only the sparse input representation, not the network structure.

6) *Performance in Adverse Weather Conditions:* We evaluate RadarGSOcc under adverse weather conditions on the unlabeled adverse-weather sequences of the K-Radar dataset (Fig. 1 and Table S-I). Since reliable per-frame ground

TABLE V
ABLATION OF RADARGSOCC WITH DIFFERENT TOP-K VALUES AND DOPPLER DIMENSIONS (SINGLE TRAINING RUNS).

Top-K	Doppler	IoU	mIoU	W-mIoU	FLOPs	FPS
150	1	35.2	6.3	20.0	1273.2	3.2
150	2	35.8	6.8	21.2	1273.2	3.2
150	3	36.5	6.8	20.4	1273.2	3.2
250	1	36.9	6.6	21.5	1273.3	3.1
250	2	34.5	6.8	20.2	1273.4	3.1
250	3	38.2	7.6	25.5	1273.5	3.1

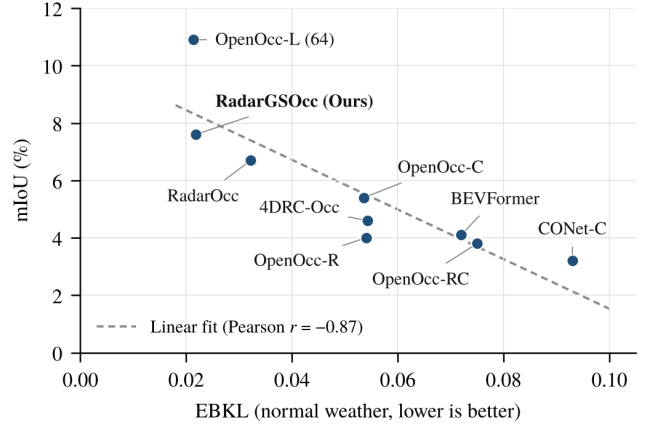


Fig. 6. mIoU versus EBKL (normal weather, test split) of the nine models in Table II for which EBKL is evaluated. The dashed line is a least-squares fit; the values are listed in Table S-IX of the supplementary material.

truth is unavailable, we use the non-empty voxel ratio (how much occupancy a model predicts) and the proposed EBKL (how close its class distribution stays to the normal-condition reference) as complementary proxies. As shown in Fig. 1 (right), RadarGSOcc attains lower EBKL than the camera- and LiDAR-based methods in rain, sleet, light snow, and heavy snow. Across the nine models evaluated under normal weather, lower EBKL goes with higher mIoU (Fig. 6; Pearson correlation coefficient -0.87), indicating that EBKL tracks mIoU. Table VI is consistent with this: LiDAR collapses under sleet, fog, and snow (e.g., 29.8% → 9.5% in fog), the camera model is stable but low, and RadarGSOcc preserves most of its occupancy through sleet, rain, and fog. In snow, however, the voxel ratio of RadarGSOcc drops to about 15% in both light and heavy snow, and under heavy snow its EBKL also doubles relative to light snow (0.05 → 0.10), which we attribute to increased 4D radar noise in dense snow [5]. Because the adverse-weather sequences also differ from the normal-weather test scenes in environment and time of day (Table S-I), both proxies reflect scene changes as well as weather. Overall, 4D radar tensors give the most consistent predictions of the three modalities, with snow as the main failure case [8].

V. CONCLUSION AND FUTURE WORK

In this work, we introduce the *K-Radar 3DSemOcc* benchmark, the first large-scale benchmark for fine-grained 3D semantic occupancy prediction using 4D radar tensors. Built on the K-Radar dataset [5], it is generated using LiDAR-based

TABLE VI

PERCENTAGE OF NON-EMPTY VOXEL PREDICTIONS UNDER DIFFERENT WEATHER CONDITIONS (LIGHT/HEAVY: LIGHT/HEAVY SNOW). THE ADVERSE-WEATHER SEQUENCES ARE LISTED IN TABLE S-I OF THE SUPPLEMENTARY MATERIAL.

Model	Normal	Sleet	Rain	Fog	Light	Heavy
OpenOcc-C [1]	18.5	17.5	16.7	16.5	17.0	18.9
CONet-L [1]	29.8	9.7	14.3	9.5	9.8	14.2
RadarGSOcc	26.5	26.1	24.5	28.1	15.1	15.3

occupancy prediction with refinement from radar annotations, covering diverse urban and highway scenes under day and night conditions. We also propose RadarGSOcc, an efficient framework that leverages geometric-semantic feature extraction and geometric refinement for 3D semantic occupancy prediction from 4D radar tensors. RadarGSOcc achieves **25.5%** weighted mIoU and **7.6%** mIoU, a **16%** relative improvement over RadarOcc [7] in weighted mIoU, while using **11% fewer parameters** and 5% fewer FLOPs. It also outperforms camera-based [1], [14] and camera–radar fusion methods, and gives more consistent predictions than camera and LiDAR under adverse weather such as rain, sleet, and snow. Its main limitations are vehicle prediction, where RadarOcc remains better when car and truck/bus are merged, and a larger run-to-run variation than RadarOcc.

Furthermore, we propose **Environment-Based KL Divergence (EBKL)** as a proxy metric to evaluate prediction consistency under adverse weather without ground truth. RadarGSOcc achieves better results in both EBKL and qualitative visualization, highlighting the effectiveness of 4D radar for robust 3D semantic occupancy prediction in adverse weather.

For future work, we aim to improve vehicle prediction and training stability, improve label quality using larger-scale datasets with accurate human annotations, and further advance robust 4D radar-based occupancy prediction in real-world environments.

ACKNOWLEDGMENT

This work was supported by the Ministry of Trade, Industry and Energy (MOTIE) grant funded by the Korea government (No. RS-2025-25451359).

REFERENCES

- [1] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, “OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 850–17 859.
- [2] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, “Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.
- [3] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, “SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21 729–21 740.
- [4] X. Huang, Z. Xu, H. Wu, J. Wang, Q. Xia, Y. Xia, J. Li, K. Gao, C. Wen, and C. Wang, “L4DR: LiDAR-4DRadar fusion for weather-robust 3D object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3806–3814.

- [5] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, “K-Radar: 4D radar object detection for autonomous driving in various weather conditions,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 3819–3829, 2022.
- [6] F. J. Abdu, Y. Zhang, M. Fu, Y. Li, and Z. Deng, “Application of deep learning on millimeter-wave radar signals: A review,” *Sensors*, vol. 21, no. 6, p. 1951, 2021.
- [7] F. Ding, X. Wen, Y. Zhu, Y. Li, and C. X. Lu, “RadarOcc: Robust 3D occupancy prediction with 4D imaging radar,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 101 589–101 617, 2024.
- [8] Y. Chae, H. Kim, and K.-J. Yoon, “Towards robust 3D object detection with LiDAR and 4D radar fusion in various weather conditions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15 162–15 172.
- [9] R. Liu, X. Wu, X. Chen, L. Hu, and Y. Lou, “4D-ROLLS: 4D radar occupancy learning via LiDAR supervision,” *arXiv preprint arXiv:2505.13905*, 2025.
- [10] L. Yang, L. Zheng, W. Ai, M. Liu, S. Li, Q. Lin, S. Yan, J. Bai, Z. Ma, and X. Zhu, “MetaOcc: Surround-view 4D radar and camera fusion framework for 3D occupancy prediction with dual training strategies,” *arXiv preprint arXiv:2501.15384*, 2025.
- [11] D. Ninfa, A. Palffy, and H. Caesar, “4DRCC: Robust semantic occupancy prediction through fusion of 4D radar and camera,” *arXiv preprint arXiv:2603.07794*, 2026.
- [12] B. Graham, “Sparse 3D convolutional neural networks,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.
- [13] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, “Vision transformer with deformable attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4794–4803.
- [14] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, “BEVFormer: Learning bird’s-eye-view representation from LiDAR-camera via spatiotemporal transformers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020–2036, 2025.
- [15] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9297–9307.
- [16] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, “K-Lane: LiDAR lane dataset and benchmark for urban roads and highways,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 4450–4459.
- [17] M. Meyer and G. Kusch, “Automotive radar dataset for deep learning based 3D object detection,” in *Proceedings of the 16th European Radar Conference (EuRAD)*, 2019, pp. 129–132.
- [18] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan *et al.*, “TJ4DRadSet: A 4D radar dataset for autonomous driving,” in *Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 493–498.
- [19] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, “Multi-class road user detection with 3+1D radar in the View-of-Delft dataset,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.
- [20] C. R. Goenawan, D.-H. Paek, and S.-H. Kong, “See the unseen: Grid-wise drivable area detection dataset and network using LiDAR,” *Remote Sensing*, vol. 16, no. 20, p. 3777, 2024.
- [21] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multi-modal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.
- [22] Y. Wang, X. Huang, X. Sun, M. Yan, S. Xing, Z. Tu, and J. Li, “UniOcc: A unified benchmark for occupancy forecasting and prediction in autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- [23] K. T. Wijaya, C. R. Goenawan, and S.-H. Kong, “Enhancing performance of 3D point completion network using consistency loss,” *Neurocomputing*, vol. 618, p. 129037, 2025.
- [24] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, “FB-OCC: 3D occupancy prediction based on forward-backward view transformation,” *arXiv preprint arXiv:2307.01492*, 2023.
- [25] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, “GaussianFormer-2: Probabilistic Gaussian superposition for efficient 3D occupancy prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 27 477–27 486.

- [26] J. Wang and S.-H. Kong, “LoD-GS: Robust and lightweight Gaussian splatting SLAM for real-time volumetric scene reconstruction,” *IEEE Robotics and Automation Letters*, vol. 11, no. 6, pp. 7668–7675, 2026.
- [27] H. Liu, Y. Chen, H. Wang, Z. Yang, T. Li, J. Zeng, L. Chen, H. Li, and L. Wang, “Fully sparse 3D occupancy prediction,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 54–71.
- [28] H. Wu, P. Lin, E. Javanmardi, N. Bao, B. Qian, H. Si, and M. Tsukada, “SUG-Occ: An explicit semantics and uncertainty guided sparse learning framework for real-time 3D occupancy prediction,” *arXiv preprint arXiv:2601.11396*, 2026.
- [29] W. Li, K. Peng, D. Wen, J. Zheng, J. Wei, M. Duan, Y. Zhang, R. Fan, and K. Yang, “Can we trust unreliable voxels? Exploring 3D semantic occupancy prediction under label noise,” *arXiv preprint arXiv:2603.06279*, 2026.
- [30] L. Chambon, E. Zablocki, A. Boulch, M. Chen, and M. Cord, “Gauss-Render: Learning 3D occupancy with Gaussian rendering,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 27 010–27 020.
- [31] B. Xu, X. Zhang, L. Wang, X. Hu, Z. Li, S. Pan, J. Li, and Y. Deng, “RPFA-Net: A 4D radar pillar feature attention network for 3D object detection,” in *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 3061–3066.
- [32] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q.-L. Han, and B. Zhu, “SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 799–812, 2024.
- [33] W. Xiong, J. Liu, T. Huang, Q.-L. Han, Y. Xia, and B. Zhu, “LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 79–92, 2024.
- [34] X. Peng, H. Sun, K. Bierzynski, A. Fischbacher, L. Servadei, and R. Wille, “MutualForce: Mutual-aware enhancement for 4D radar-LiDAR 3D object detection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [35] S.-H. Kong, D.-H. Paek, and S. Lee, “RTNH+: Enhanced 4D radar object detection network using two-level preprocessing and vertical encoding,” *IEEE Transactions on Intelligent Vehicles*, vol. 10, pp. 1427–1440, 2025.
- [36] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, “VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9087–9098.
- [37] L. Zheng, J. Liu, R. Guan, L. Yang, S. Lu, Y. Li, X. Bai, J. Bai, Z. Ma, H.-L. Shen, and X. Zhu, “Doracamom: Joint 3D detection and occupancy prediction with multi-view 4D radars and cameras for omnidirectional perception,” *arXiv preprint arXiv:2501.15394*, 2025.
- [38] Y. Ma, J. Mei, X. Yang, L. Wen, W. Xu, J. Zhang, X. Zuo, B. Shi, and Y. Liu, “LiCROcc: Teach radar for accurate semantic occupancy prediction using LiDAR and camera,” *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 852–859, 2025.
- [39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.
- [40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.
- [42] M. Berman, A. R. Triki, and M. B. Blaschko, “The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4413–4421.
- [43] A.-Q. Cao and R. de Charette, “MonoScene: Monocular 3D semantic scene completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3991–4001.
- [44] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception

for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2446–2454.



Christofel Rio Goenawan received the B.S. degree in electrical engineering from the Bandung Institute of Technology (ITB), Indonesia, in 2021. He is currently pursuing the M.S. degree with the Robotics Program, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He has three years of industry experience as a data scientist and machine learning engineer in the telecommunication, FMCG, banking, and marketing sectors. His current research interests include perception and end-to-end autonomous driving using AI.



Dong-Hee Paek (Member, IEEE) is currently an Assistant Professor with the Department of Future Mobility, Korea University, Sejong, Korea. He received the B.S. degree in robotics and intelligent systems from Kwangju University, Seoul, Korea, in 2019, and the M.S. degree in robotics and the Ph.D. degree in mobility and robotics from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2021 and 2025, respectively. From 2024 to 2025, he was with Zeta Mobility, Daejeon, Korea, as a Senior Research Engineer, where he conducted research and development on 4D radar perception and sensor fusion AI while pursuing the Ph.D. degree. From 2025 to 2026, he was a Postdoctoral Research Fellow with KAIST. His current research interests include 4D radar signal processing and perception, multi-modal sensor fusion, and AI-based robotic perception.



Seung-Hyun Kong is an Associate Professor in the CCS Graduate School of Mobility, Korea Advanced Institute of Science and Technology (KAIST), where he has been a faculty member since 2010. He received a B.S. degree in Electronics Engineering from Sogang University, Seoul, Korea, in 1992, an M.S. degree in Electrical and Computer Engineering from Polytechnic University (now NYU), New York, in 1994, and a Ph.D. degree in Aeronautics and Astronautics from Stanford University, Palo Alto, in 2005. From 1997 to 2004 and from 2006 to 2010, he was with companies, including Samsung Electronics (Telecommunication Research Center) and Nexpilot, both in Korea, Polaris Wireless in San Jose, and Qualcomm (Corporate R&D) in San Diego, for advanced R&D in mobile communication systems, wireless positioning, and assisted GNSS. His current research interest includes deep neural networks for perception and localization using 4D radar and sensor fusion, reinforcement and imitation learning algorithms for end-to-end autonomous driving, and neural network compression for embedded systems. He has authored more than 100 papers in peer-reviewed journals and conference proceedings and 14 patents, and his research group won the President’s Award in the 2018 International Student Autonomous Driving Competition from the Korean government. He served as a Program co-Chair of IEEE ITSC 2019, New Zealand, and the General Chair of IEEE IV 2024, Korea, and he has served as an Associate Editor of IEEE T-ITS since 2017.

Supplementary Material

RadarGSOcc: Doppler-Aware 3D Semantic Occupancy Prediction Using 4D Radar Tensor with Geometric-Semantic Feature Extraction

TABLE S-I

COMPOSITION OF THE K-RADAR 3DSEMOCC BENCHMARK AND THE UNLABELED ADVERSE-WEATHER SEQUENCES.

Scene	Split	Weather	Environment	Time	#Frames
2	Train	Normal	Highway	Night	462
5	Train	Normal	Urban	Day	595
8	Train	Normal	University	Night	567
9	Train	Normal	Highway	Day	833
10	Train	Normal	Highway	Day	1,130
14	Train	Normal	Urban	Day	595
16	Train	Normal	University	Day	578
17	Train	Normal	University	Day	593
20	Train	Normal	Urban	Day	557
Training total					5,910
3	Test	Normal	Highway	Night	596
6	Test	Normal	Urban	Night	594
15	Test	Normal	Urban	Day	591
Test total					1,781
23	Adverse	Rain	Urban	Night	598
24	Adverse	Rain	Urban	Night	598
27	Adverse	Sleet	Suburban	Day	597
34	Adverse	Rain	Suburban	Night	598
38	Adverse	Fog	Mountain	Day	597
46	Adverse	Heavy snow	Highway	Night	579
48	Adverse	Light snow	Highway	Night	569
50	Adverse	Sleet	Highway	Night	590
51	Adverse	Sleet	Highway	Night	589
54	Adverse	Heavy snow	Urban	Day	573
55	Adverse	Heavy snow	Urban	Day	572

This supplementary material provides details of the K-Radar 3DSemOcc benchmark (scene composition, class statistics, and label verification), the definition of the Doppler mean/variance channels, implementation details of the deformable attention modules, results for all semantic classes, the environment-wise class-distribution statistics that motivate the EBKL proxy metric, per-run results, and the data behind the mIoU-EBKL correlation. Sections, tables, and figures are prefixed with “S” to distinguish them from those of the main paper; references are listed in the main paper.

S-I. K-RADAR 3DSEMOCC BENCHMARK DETAILS

A. Scene Composition

The benchmark is built from 12 K-Radar sequences: scenes 2, 5, 8, 9, 10, 14, 16, 17, and 20 for training (5,910 frames) and scenes 3, 6, and 15 for testing (1,781 frames). All labeled scenes are recorded under normal weather conditions. Table S-I lists the environment, time of day, and number of frames of each scene, together with the unlabeled adverse-weather sequences used in Section IV-D6 of the main paper, which are disjoint from the labeled scenes. EBKL is not reported for the fog sequence (scene 38), which is used only in the voxel-ratio analysis (Table VI).

TABLE S-II

GROUND-TRUTH VOXEL STATISTICS PER CLASS BEFORE LABEL REFINEMENT (LABEL-CONFIDENCE THRESHOLD 0.8).

Class	#Voxels (train)	%	#Voxels (test)	%
Drivable area	4,183,019	24.12	1,216,791	26.67
Vegetation	8,536,616	49.22	2,068,116	45.34
Manmade	1,776,829	10.24	382,270	8.38
Terrain	2,056,773	11.86	710,776	15.58
Sidewalk	406,188	2.34	91,606	2.01
Other flat	12,819	0.07	6,723	0.15
Barrier	5,678	0.03	1,629	0.04
Car	161,384	0.93	39,778	0.87
Truck/bus	171,656	0.99	36,146	0.79
Trailer	17,235	0.10	3,111	0.07
Traffic cone	1,574	0.01	710	0.02
Pedestrian	12,218	0.07	3,818	0.08
Motorcycle	2,195	0.01	145	<0.01
Bicycle	209	<0.01	35	<0.01
Total	17,344,393	100	4,561,654	100

TABLE S-III

GROUND-TRUTH VOXEL STATISTICS PER CLASS AFTER LABEL REFINEMENT (LABEL-CONFIDENCE THRESHOLD 0.8), USED IN ALL EXPERIMENTS.

Class	#Voxels (train)	%	#Voxels (test)	%
Drivable area	3,738,578	22.94	1,206,770	25.95
Vegetation	7,977,664	48.96	2,048,675	44.06
Manmade	1,526,764	9.37	372,645	8.01
Terrain	1,888,009	11.59	704,612	15.15
Sidewalk	378,405	2.32	91,151	1.96
Other flat	12,065	0.07	6,518	0.14
Barrier	4,828	0.03	1,477	0.03
Car	412,504	2.53	75,112	1.62
Truck/bus	318,448	1.95	137,398	2.95
Trailer	10,495	0.06	1,203	0.03
Traffic cone	1,324	0.01	702	0.02
Pedestrian	14,216	0.09	2,867	0.06
Motorcycle	3,593	0.02	699	0.02
Bicycle	8,262	0.05	35	<0.01
Total	16,295,155	100	4,649,864	100

B. Class Statistics

Tables S-II and S-III report the number and percentage of ground-truth voxels per class in the training and test splits before and after the label refinement of Section III-A; all experiments use the refined labels. Small foreground classes (traffic cone, pedestrian, motorcycle, and bicycle) each account for less than 0.1% of all voxels, which explains the near-zero IoU of all methods on these classes and motivates the use of W-mIoU in the main paper. Refinement increases the number of voxels of the moving-object classes (car, truck/bus, trailer, pedestrian, motorcycle, and bicycle) in the training split from 364,897 to 767,518.

The benchmark therefore has 15 classes in total: the 14 semantic classes listed in Table S-III and the empty class. All IoU-based metrics in the main paper are computed over the

TABLE S-IV
MANUAL VERIFICATION OF 525 RANDOMLY SAMPLED NON-EMPTY VOXELS.

Class	#Sampled	#Correct	Accuracy (%)
Drivable area	126	125	99.2
Vegetation	191	188	98.4
Manmade	66	64	97.0
Terrain	99	69	69.7
Car	16	14	87.5
Truck/bus	5	4	80.0
Other classes	22	18	81.8
All	525	482	91.8

14 semantic classes.

C. Label Verification

To assess the quality of the refined pseudo-labels, 525 non-empty voxels were randomly sampled from the benchmark (label-confidence threshold 0.8) and verified manually. Each sampled voxel is drawn as a green 3D box and projected onto the front-camera image and onto the bird’s-eye view of the LiDAR point cloud of the same frame (Fig. S1); a voxel is counted as correct when its semantic label agrees with the object or surface visible at that location in both views. The overall accuracy is 91.8% (482 of 525), and per-class results are given in Table S-IV; terrain has the lowest accuracy (69.7%).

S-II. DOPPLER MEAN AND VARIANCE CHANNELS

In addition to the top-3 power values and their Doppler indices, the 8-channel sparse tensor contains the mean and the variance of the power along the Doppler dimension of each spatial element, following RadarOcc [7]. For a spatial element with Doppler bins $d = 1, \dots, N_D$ and corresponding power p_d , they are defined as

$$\mu_D = \frac{1}{N_D} \sum_{d=1}^{N_D} p_d, \quad (20)$$

$$\sigma_D^2 = \frac{1}{N_D} \sum_{d=1}^{N_D} (p_d - \mu_D)^2. \quad (21)$$

S-III. IMPLEMENTATION DETAILS OF THE DEFORMABLE ATTENTION MODULES

RadarGSOcc uses 3D deformable attention in three places: the deformable self-attention in the geometric and semantic branches, the geometric-semantic cross-attention fusion, and the geometric feature refinement in the occupancy head. Table S-V summarizes their hyperparameters.

The geometric and semantic branches use a feature dimension of 96 each, so that their concatenation matches the 192-dimensional occupancy features used in the refinement module and in RadarOcc. Following RadarOcc, the reference points are initialized uniformly along the range, azimuth, and elevation dimensions of the spherical grid, and the sampling locations $ref + \Delta$ are expressed in normalized coordinates in $[0, 1]$ along each axis (0 and 1 denote the two ends of the axis).

TABLE S-V
HYPERPARAMETERS OF THE DEFORMABLE ATTENTION MODULES.

	Self-attn.	Fusion	Refinement
Feature dimension dim	96	96	192
Number of heads	8	8	8
Sampling points per query	8	8	8
Number of layers	1	1	1
Sampling coordinates	$[0, 1]$	$[0, 1]$	$[0, 1]$

Features at the sampled locations are obtained by trilinear interpolation, and the softmax in the attention equations of Section III-B2 and Section III-B3 of the main paper is taken over the sampled locations.

S-IV. RESULTS FOR ALL SEMANTIC CLASSES

Table S-VI extends Table II of the main paper to all semantic classes. Except for OpenOcc-C, the camera-based, radar point cloud-based, and camera–radar fusion baselines obtain zero IoU on all seven classes on the right.

S-V. ENVIRONMENT-WISE CLASS DISTRIBUTION FOR EBKL

EBKL assumes that, within the same environment type (urban, university campus, or highway), the class distribution of the scene is relatively stable across scenes and weather conditions. Fig. S2 shows the normalized ground-truth class distribution of the normal-weather scenes grouped by environment, and Table S-VII reports the mean EBKL between the individual scenes of each environment (rows) and each environment-level reference distribution (columns). Highway scenes are clearly closer to the highway reference (0.015) than to the urban and university references (0.032 and 0.036), and university scenes are closest to their own reference (0.008). Urban scenes, however, are slightly closer to the university reference (0.014) than to their own (0.017), so the urban and university-campus distributions are not well separated; in K-Radar, EBKL therefore mainly distinguishes highway from non-highway scenes.

S-VI. PER-RUN RESULTS

Table S-VIII lists the results of the three training runs of RadarOcc and RadarGSOcc that are summarized in Section IV-D2 of the main paper. The first run of each model was trained without a fixed random seed and is the run reported in Table II and the other tables of the main paper; the second and third runs use seeds 1111 and 2222. The means and standard deviations are computed from the unrounded values.

S-VII. DATA BEHIND THE mIoU–EBKL CORRELATION

Table S-IX lists the mIoU (Table II) and the EBKL under normal weather of the nine models used for Fig. 6 of the main paper. Over these nine models, the Pearson correlation coefficient between mIoU and EBKL is -0.87 (Spearman rank correlation -0.95); models with higher mIoU have lower EBKL.

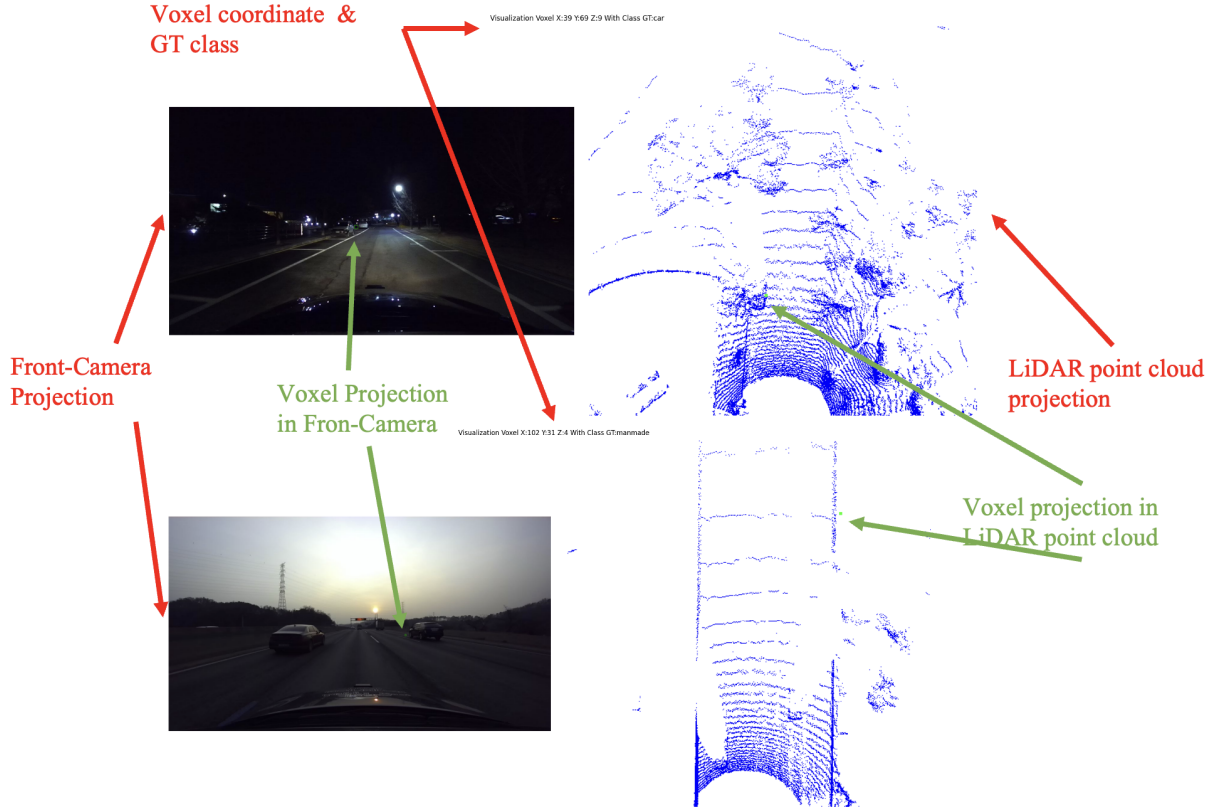


Fig. S1. Manual verification of the K-Radar 3DSemOcc benchmark. Sampled voxels (green 3D boxes) are projected onto the front-camera image (left) and the LiDAR bird's-eye view (right): (top) car voxels in a night scene and (bottom) manmade voxels in a day scene.

TABLE S-VI
3D SEMANTIC OCCUPANCY RESULTS ON THE K-RADAR 3DSEMOC benchmark FOR ALL CLASSES (IoU in %, TEST SPLIT).

Method	Sensor	IoU	mIoU	W-mIoU	DA	Veg	Man	Car	Truck	Ter	Side	Other	Trail	Cone	Ped	Moto	Bike	Bar
CONet-C [1]	Cam	16.9	3.2	11.3	27.4	4.2	0.0	0.6	0.0	11.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BEVFormer [14]	Cam	27.3	4.1	16.0	34.0	10.4	0.1	0.0	0.0	11.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OpenOcc-C [1]	Cam	30.0	5.4	17.1	34.7	17.3	13.7	5.9	1.8	19.0	0.9	1.6	0.0	0.0	0.0	0.0	0.0	1.5
OpenOcc-R [1]	Radar	26.0	4.0	14.7	34.0	8.8	5.1	1.2	0.0	8.9	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OpenOcc-RC [1]	R+Cam	25.2	3.8	14.0	34.9	5.6	0.8	0.0	0.1	11.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4DRC-Occ [11]	R+Cam	30.2	4.6	17.6	35.2	12.8	0.1	0.2	0.0	15.3	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OpenOcc-L (64) [1]	LiDAR	53.1	10.9	36.1	50.5	39.9	15.4	3.3	2.4	26.1	9.1	6.4	0.0	0.0	0.2	0.0	0.0	0.0
OpenOcc-L (128) [1]	LiDAR	54.3	11.1	38.9	51.5	43.0	15.4	1.5	3.3	30.3	6.4	3.8	0.2	0.0	0.2	0.0	0.0	0.8
RadarOcc [7]	RadarT	36.4	6.7	21.9	40.4	16.3	10.3	6.3	0.6	18.0	2.4	0.7	0.0	0.0	0.1	0.0	0.0	0.5
RadarGSOcc (Ours)	RadarT	38.2	7.6	25.5	42.2	21.2	11.6	0.6	4.7	23.9	2.5	0.2	0.0	0.0	0.0	0.0	0.0	0.6
+ w/o overlap	RadarT	35.7	6.8	21.8	34.5	17.5	11.9	0.6	2.1	21.2	4.9	0.1	1.3	0.0	0.8	0.2	0.0	0.4

TABLE S-VII
MEAN EBKL BETWEEN SCENE-LEVEL AND ENVIRONMENT-LEVEL CLASS DISTRIBUTIONS (GROUND TRUTH, NORMAL WEATHER).

Scenes \ Reference	Urban	University	Highway
Urban scenes	0.017	0.014	0.030
University scenes	0.015	0.008	0.035
Highway scenes	0.032	0.036	0.015

TABLE S-VIII
RESULTS OF THE THREE TRAINING RUNS OF RADAROC and RADARGSOCC ON THE K-RADAR 3DSEMOC benchmark (TEST SPLIT).

Model	Run (seed)	IoU	mIoU	W-mIoU
RadarOcc	1 (not fixed)	36.4	6.7	21.9
	2 (1111)	36.3	6.4	20.1
	3 (2222)	36.0	7.0	20.8
	Mean $\pm$ std	36.26 $\pm$ 0.23	6.72 $\pm$ 0.28	20.93 $\pm$ 0.90
RadarGSOcc	1 (not fixed)	38.2	7.6	25.5
	2 (1111)	39.3	7.0	24.6
	3 (2222)	34.8	6.8	21.4
	Mean $\pm$ std	37.43 $\pm$ 2.34	7.14 $\pm$ 0.41	23.83 $\pm$ 2.15

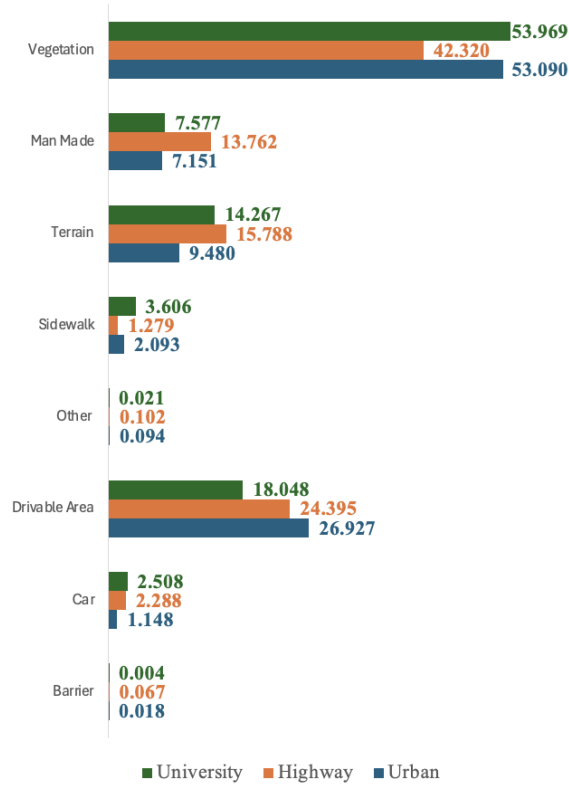


Fig. S2. Normalized ground-truth class distributions (%) of the normal-weather scenes, grouped by environment, for the eight classes shown.

TABLE S-IX
MIOU AND EBKL (NORMAL WEATHER, TEST SPLIT) OF THE NINE
MODELS IN FIG. 6.

Model	mIoU (%)	EBKL
CONet-C [1]	3.2	0.093
BEVFormer [14]	4.1	0.072
OpenOcc-C [1]	5.4	0.0536
OpenOcc-R [1]	4.0	0.054
OpenOcc-RC [1]	3.8	0.075
4DRC-Occ [11]	4.6	0.0543
OpenOcc-L (64) [1]	10.9	0.0214
RadarOcc [7]	6.7	0.0322
RadarGSOcc (Ours)	7.6	0.0218