







LiCROcc: Teach Radar for Accurate Semantic Occupancy Prediction Using LiDAR and Camera

Yukai Ma , Jianbiao Mei , Xuemeng Yang, Licheng Wen , Weihua Xu, Jiangning Zhang , Xingxing Zuo , Botian Shi , and Yong Liu 

Abstract—Semantic Scene Completion (SSC) is pivotal in autonomous driving perception, frequently confronted with the complexities of weather and illumination changes. The long-term strategy involves fusing multi-modal information to bolster the system’s robustness. Radar, increasingly utilized for 3D target detection, is gradually replacing LiDAR in autonomous driving applications, offering a robust sensing alternative. In this letter, we focus on the potential of 3D radar in semantic scene completion, pioneering cross-modal refinement techniques for improved robustness against weather and illumination changes and enhancing SSC performance. Regarding model architecture, we propose a three-stage tight fusion approach on BEV to realize a fusion framework for point clouds and images. Based on this foundation, we designed three cross-modal distillation modules—CMRD, BRD, and PDD. Our approach enhances the performance in radar-only (R-LiCROcc) and radar-camera (RC-LiCROcc) settings by distilling to them the rich semantic and structural information of the fused features of LiDAR and camera. Finally, our LC-Fusion, R-LiCROcc and RC-LiCROcc achieve the best performance on the nuScenes-Occupancy dataset, with mIOU exceeding the baseline by 22.9%, 44.1%, and 15.5%, respectively.

Index Terms—Sensor fusion, semantic scene completion, knowledge distillation.

I. INTRODUCTION

SEMANtic Scene Completion (SSC), a crucial technology in autonomous driving, has garnered substantial attention for its ability to ground detailed 3D scene information. Cameras and LiDAR are the most prevalent sensors used for SSC tasks, each with strengths and limitations. The former provides rich semantic context but lacks depth information and is susceptible

to lighting and weather conditions. The latter offers accurate 3D geometry but performs poorly when given highly sparse input and is hindered for wide applications due to the high cost of dense LiDAR sensors. On the other hand, radar, a weather-resistant sensor gaining traction in autonomous driving, is valued for its automotive-grade design and affordability. Despite its robustness in diverse weather and lighting conditions [1], [2], radar’s sparse and noisy measurements present significant challenges for SSC in large-scale outdoor scenarios. Consequently, we pose the question: Can we devise a radar-centric method that achieves satisfactory SSC performance?

Most research has recently focused mainly on radar-based detection [6], [7]. Only a few studies [1], [8] have explored the application of radar sensors in the SSC task. However, RadarOcc [8] can only use radar to predict occupancy in very few categories or as a supplement to multi-modal inputs. The focus of OccFusion [1] is on the module of multi-modal fusion, where radar, as an input to the modality, is only a part of the system. Instead, we use radar as the primary sensor to explore its performance on SSC tasks containing more categories. In addition, we found that although the radar has inherent strengths against adverse weather conditions and illumination changes, as indicated in Table I and Fig. 1, there is still a significant performance gap between the radar-based and LiDAR/camera-based SSC approaches. [9], [10] employ knowledge distillation (KD) to improve radar performance in target detection. However, none of these methods apply to this study due to differences in tasks and the design of the distillation module, which is exclusively intended for distillation involving the same modal combinations, such as LiDAR-camera to radar-camera or LiDAR to radar.

To address these challenges, we propose a radar-centric SSC network, with the camera as an optional supplementary aid. Additionally, we have designed new cross-modal distillation modules to accommodate various combinations of sensors. As illustrated in Fig. 1, we first design LC-Fusion, the most comprehensive LiDAR-Camera fusion network, to serve as our teacher model. RC-Fusion follows the architecture of LC-Fusion and supports a combination of radar and camera inputs or radar-only configurations. For the fusion-based KD module, we combine Cross-Model Residual Distillation (CMRD), BEV Relation Distillation (BRD), and Predictive Distribution Distillation (PPD) to hierarchically compel the student model to learn the feature representations and distributions of the teacher model. While maintaining robustness in adverse weather conditions and night vision capabilities, our LiCROcc with radar (R-LiCROcc) achieves comparable results against camera-based methods, and LiCROcc with radar and camera (RC-LiCROcc) approaches the

Received 21 July 2024; accepted 6 November 2024. Date of publication 4 December 2024; date of current version 17 December 2024. This article was recommended for publication by Editor A. Valada upon evaluation of the Associate Editor and reviewers’ comments. This work was supported by the National Natural Science Foundation of China under Grant U21A20484. (Yukai Ma and Jianbiao Mei are co-first authors.) (Corresponding authors: Botian Shi; Yong Liu.)

Yukai Ma and Jianbiao Mei are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Xuemeng Yang, Licheng Wen, and Botian Shi are with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: shibotian@pjlab.org.cn).

Weihua Xu, Jiangning Zhang, and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: yongliu@iipc.zju.edu.cn).

Xingxing Zuo is with the School of Computation, Information and Technology, Technical University of Munich, 80333 München, Germany.

The project page is available at <https://hr-zju.github.io/LiCROcc/>.

Digital Object Identifier 10.1109/LRA.2024.3511427

TABLE I

PERFORMANCE ON NUSCENES-OCCUPANCY (VALIDATION SET). WE REPORT THE GEOMETRIC METRIC IOU, SEMANTIC METRIC mIOU, AND THE IOU FOR EACH SEMANTIC CLASS. THE *C, D, L, R, M* DENOTES CAMERA, DEPTH, LiDAR, RADAR AND MULTI-MODAL. FOR *SURROUND*= \checkmark , THE METHOD DIRECTLY PREDICTS SURROUNDING SEMANTIC OCCUPANCY WITH 360-DEGREE INPUTS. OTHERWISE, THE METHOD PRODUCES THE RESULTS OF EACH CAMERA VIEW, AND THEN CONCATENATES THEM AS SURROUNDING OUTPUTS. WE DIVIDE THE FORM INTO THREE CATEGORIES BASED ON THE MODALITY OF THE INPUTS. BOLD REPRESENTS THE BEST SCORE

Method	Input Surround		IoU mIoU		barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
MonoScene [23]	C	✗	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [20]	C	✓	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [33]	C&D	✗	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [34]	C&D	✗	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSCNet [35]	L	✓	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [36]	L	✓	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
C-CONet [3]	C	✓	20.1	12.8	13.2	8.1	15.4	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8	18.7	16.3	4.8	8.2
L-CONet [3]	L	✓	30.9	15.8	17.5	5.2	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5	22.4	21.7	19.2	23.5
C-LiCROcc (Ours)	C	✓	23.2	13.5	15.0	9.0	13.7	17.5	7.6	10.4	11.5	8.4	6.5	12.2	31.5	18.9	19.3	17.9	6.3	10.6
PointOcc [5]	L	✓	34.1	23.9	24.9	19.0	20.9	25.7	13.4	25.6	30.6	17.9	16.7	21.2	36.5	25.6	25.7	24.9	24.8	29.0
M-CONet [3]	C&L	✓	29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ [4]	C&L	✓	30.6	21.9	26.5	16.8	22.3	27.0	10.1	20.9	20.7	14.5	16.4	21.6	36.9	23.5	25.5	23.7	20.5	23.5
LC-Fusion (Ours)	C&L	✓	34.9	24.7	29.6	20.5	22.2	26.4	15.7	24.5	27.3	21.8	18.1	21.8	35.9	22.8	25.0	25.1	27.8	30.5
R-CONet [3]	R	✓	17.0	5.9	6.3	0.6	3.4	9.4	0.9	0.9	1.0	1.7	2.3	3.9	24.2	8.8	11.4	8.6	6.1	4.2
R-SSC-RS [15]	R	✓	20.8	7.5	5.3	0.3	5.4	13.1	1.7	1.4	7.4	2.3	2.6	7.0	24.2	8.5	11.4	9.4	9.2	10.5
PointOcc [5]	R	✓	21.9	7.4	4.9	0.8	5.7	13.1	1.6	2.1	6.1	1.6	2.5	5.9	26.4	8.1	11.7	8.7	9.9	10.0
R-LiCROcc (Ours)	R	✓	21.3	8.5	8.1	0.9	6.4	13.6	2.3	2.7	7.9	2.7	3.3	7.6	24.3	11.2	13.1	10.7	10.6	11.0
RC-CONet [3]	C&R	✓	19.0	14.8	16.7	11.6	17.0	20.6	8.6	15.5	15.0	11.2	6.5	15.5	28.0	19.7	18.7	15.6	7.7	8.8
Co-Occ [4]	C&R	✓	24.2	16.6	18.6	12.6	18.1	23.0	6.4	16.5	15.2	11.2	7.0	15.3	34.3	21.9	23.0	19.6	10.3	12.0
RC-Fusion (Ours)	C&R	✓	25.2	15.6	14.6	10.4	16.4	20.5	9.5	15.0	15.5	10.0	7.0	15.0	32.2	18.5	20.3	18.3	11.6	14.8
RC-LiCROcc (Ours)	C&R	✓	26.0	17.1	18.6	11.9	17.1	21.6	11.1	15.5	16.7	11.5	8.8	16.0	34.1	20.9	21.9	19.7	12.8	15.9

performance of LiDAR-based methods. To summarize, the main contributions are as follows:

- We aim to improve radar for semantic scene completion while preserving real-world practicality, leveraging radar's resilience to various weather conditions. We also establish radar-based benchmarks from LiDAR-based approaches, fostering radar-based SSC research, and consider a camera-radar fusion network for enhanced performance.
- We resent a new radar-centric framework, LiCROcc, which combines CMRD, BRD, and PPD modules to hierarchically force the student model to learn the feature representations and distributions of the teacher model.
- Extensive experiments on the large-scale nuScenes-Occupancy [3] demonstrate the effectiveness of our proposed approaches.

II. RELATED WORK

A. Radar for Segmentation and 3D Object Detection

The initial radar use for detection and segmentation often involved integration with other modules due to sparse point cloud characteristics. [6], [11] converts image features into BEV using radar points, integrating maps with a multi-modal attention mechanism. LXL [12] merges radar and image features in the BEV perspective. The fusion of 4D radar and LiDAR has received less attention due to similar point cloud forms. InterFusion [13] facilitates information exchange between 4D radar and LiDAR at the pillar level, mitigating information loss. Conversely, using adaptive weights, RLNet [14] achieves radar

and LiDAR fusion at the voxel level. These approaches primarily utilize radar to produce sparse perception results.

B. 3D Semantic Scene Completion

LiDAR/Camera-based methods: LiDAR-based methods [5], [15], [16], [17] use LiDARs for precise 3D semantic occupancy prediction.

Advanced methods focus on multi-view fusion [18], local implicit functions [19], knowledge distillation [16], and BEV representation [15]. Recently, Pasco [17] further extends the SSC task with instance-level information to produce a richer 3D scene understanding. Camera-based methods [20], [21], [22], [23] have become popular due to their rich visual cues and cost-effectiveness. Many methods have explored effective 3D scene representation learning for outdoor surrounding SSC. For instance, TPVFormer [20] introduces a tri-perspective view for detailed 3D structure representation. OccFormer [21] utilizes transformers for extracting multi-scale voxel features.

Multi-modal methods [1], [3], [4] combines multi-source sensor data (e.g., images, LiDARs, and radars) to perform robust outdoor SSC. OpenOccupancy [3] provides a surrounding benchmark and establishes camera-based, LiDAR-based and LiDAR-camera baselines. Recently, radar perception [24], [25], [26] has garnered wide attention in multi-modal 3D detection task. However, there are only a few works [1] to incorporate radar for outdoor SSC tasks. For instance, the recent OccFusion [1] devises a sensor fusion framework to integrate features from LiDARs, surround view images, and radars for robust and

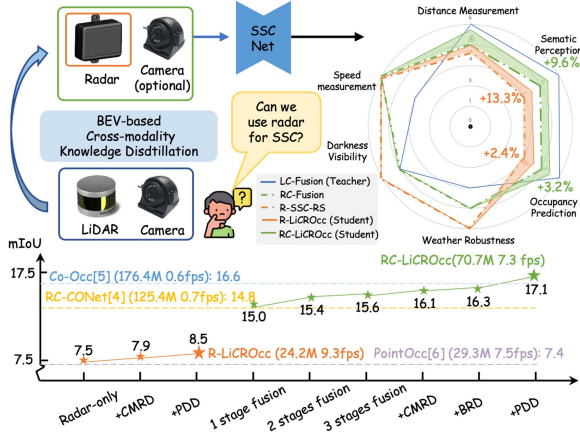


Fig. 1. Our motivation for proposing LiCROcc. We explore radar performance on SSC tasks with the expectation of developing a network with balanced performance and robustness. We further improve radar sensors' semantic occupancy prediction and distance measurement capabilities by cross-modal distillation while maintaining their inherent night vision capabilities and weather robustness. In the figure on the right, the blue line represents the capability of the teacher model, the dashed line represents the capability of the pre-distillation student in our two settings, and the solid line in the same color represents the equilibrium performance of the model after KD. The graphs on the bottom demonstrate the enhancement of our method after gradually adding modules and the comparison with other methods on nuScenes-Occupancy [3] (validation set), including Co-Occ [4], RC-CoNet [3], and PointOcc [5].

accurate SSC. Our radar-centric framework focuses more on the distillation design of LC modes distilled to RC or R than on the direct fusion of multiple modes.

C. Knowledge Distillation in Semantic Scene Completion

Knowledge Distillation (KD) was initially proposed for model compression and performance improvement in image classification tasks [27]. It has since been extended to other fields, such as 3D object detection [24], [28], [29], 3D segmentation [30], [31], and SSC [16], [32].

In SSC, SCPNet [16] uses Dense-to-Sparse KD (DSKD) to transfer dense, relation-based semantic knowledge from a multi-LiDAR teacher to a single-LiDAR student, boosting the student's representation learning. Similarly, MonoOcc [32] employs KD to transfer the temporal information to a monocular semantic occupancy framework. In contrast, our LiCROcc introduces cross-modal knowledge distillation in a shared BEV space for outdoor SSC.

III. METHODOLOGY

A. Overview

As mentioned above, we construct the radar-centric baseline and design the radar-camera fusion network (the bottom part of Fig. 2) to boost the baseline's performance. To leverage the guidance of detailed geometric structure and point representation in LiDAR-camera fusion, we further utilize the fusion-based KD (Section III-C) to transfer the knowledge from the LiDAR-camera fusion network (the top part of Fig. 2) to the radar-based baseline and radar-camera fusion network. We employ the same architecture, i.e., the multi-modal fusion network (Section III-B), to establish the above two fusion networks.

B. Multi-Modal Fusion Network

The multi-modal fusion network mainly consists of the image branch for extracting image features, the point branch for encoding LiDAR/radar points, and the multi-modal BEV fusion network for effectively and efficiently fusing point and image features hierarchically.

Image branch: Following FlashOcc [22], we propose to project surrounding image features to BEV space for subsequent processing, alleviating the memory overhead while maintaining the high accuracy of occupancy prediction. As shown in Fig. 2, the image branch mainly consists of three components: the camera encoder for image features, the Perspective View to BEV projection layer for BEV representation of the 3D scene, and the BEV encoder for hierarchical BEV features ($F_{c,0}, F_{c,1}, F_{c,2}$, where $F_{c,i} \in \mathbb{R}^{C_i \times H_i \times W_i}$) that contain the rich semantic context. The extracted multi-scale BEV features are fed into the BEV fusion model to interact with the point features, which will be elaborated on below.

Point branch: Without losing generality, we adopt the recent BEV-based SSC-RS [15] as our point branch. This branch uses two independent branches for semantic and geometric encoding. The BEV fusion network with an ARF module [15] aggregates features from these branches, resulting in the final SSC. Due to its disentangled design, SSC-RS is lightweight and has strong representation ability, making it very suitable for use as the point branch. The point branch takes the LiDAR/radar point cloud P and outputs the multi-scale BEV features ($F_{p,0}, F_{p,1}, F_{p,2}$, where $F_{p,i} \in \mathbb{R}^{C_i \times H_i \times W_i}$). For LiDAR point cloud, the $P \in \mathbb{R}^{N \times 4}$ is in the range of $[R_x, R_y, R_z, R_{\text{intensity}}]$. Moreover, for the radar point cloud, $P \in \mathbb{R}^{N \times 7}$ is the concatenation of the xyz coordinates, radar cross section σ , the xy velocities compensated by the ego-motion and the dummy field for sweep info.

Multi-modal BEV Fusion network: Due to the computational burden of 3D convolutions for dense feature fusion, we introduce a multi-modal BEV fusion network, drawing inspiration from BEV perception tasks. This network efficiently combines semantically rich visual BEV representations ($F_{c,0}, F_{c,1}, F_{c,2}$), geometrically informative LiDAR features or weather-resistant radar features. To streamline the fusion process, we unify LiDAR or radar point cloud features with ($F_{p,0}, F_{p,1}, F_{p,2}$). Similarly to [15], our BEV fusion network employs a 2D convolutional U-Net architecture. Each residual block reduces the input feature resolution by a factor of 2 to maintain consistency with semantic/complementary features. Before each subsequent block, we integrate the previous stage's $F_{b,i-1}$ with the current stage's $F_{p,i}$ using ARF [15] to obtain $F_{b,i}$, and then the scaled $F_{c,i}$ is fused to $F_{b,i}$ by addition. The decoder upsamples the encoder's compressed features three times by a factor of two by skipping connections. The final decoder convolution generates the SSC prediction $Y \in \mathbb{R}^{((C_n+1) \cdot Z) \times H \times W}$, where C_n denotes the number of semantic classes. To represent the voxel-wise semantic occupancy probabilities, Y is reshaped into $((C_n+1) \times Z \times H \times W)$. To train the proposed fusion model, cross-entropy loss \mathcal{L}_{ce} is used to optimize the network. In addition, following [23], we also utilize affinity loss $\mathcal{L}_{\text{scal}}^{\text{geo}}$ and $\mathcal{L}_{\text{scal}}^{\text{sem}}$ to optimize the metrics in the scene and the class (i.e., geometric IoU, and semantic mIoU). Therefore, the BEV fusion loss function can be derived as:

$$\mathcal{L}_{\text{ss}} = \mathcal{L}_{ce} + \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{scal}}^{\text{sem}}, \quad (1)$$

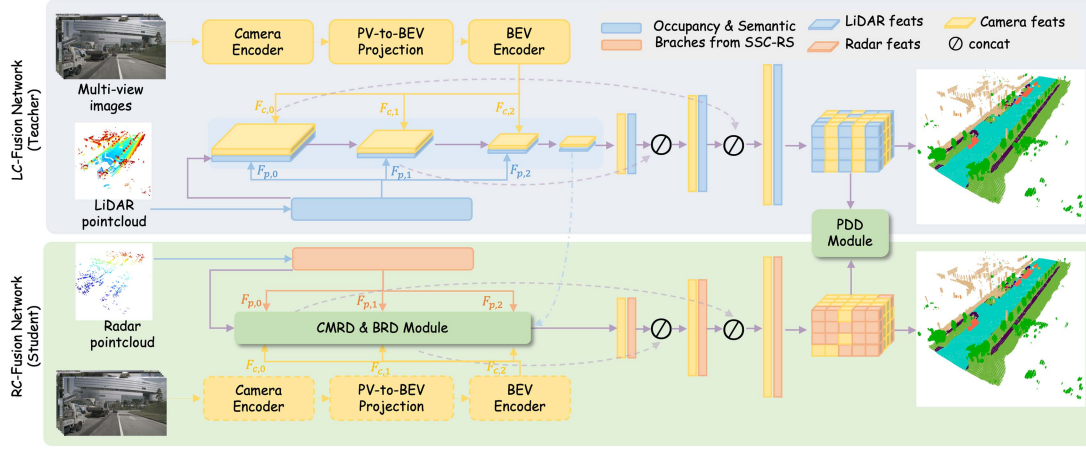


Fig. 2. Overall framework of our LiCROcc. We designed a base framework for point cloud and image fusion, where the models serve as the teacher (LC-Fusion Network) and student (RC-Fusion Network), respectively. We unified both fusion processes under BEV space to reduce computational cost. Additionally, we designed three novel distillation losses (CMRD, BRD, and PDD) to achieve effective cross-modal KD. Inference is performed using only the RC fusion Network, where camera input is optional. For the detailed structure of CMRD and BRD, please refer to Fig. 3.

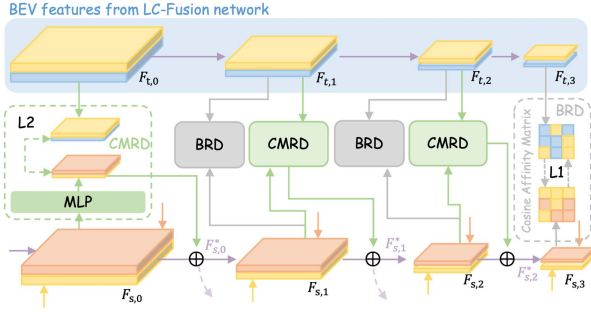


Fig. 3. Detailed illustration of the use of CMRD and BRD on BEV features. The feature with the blue background is a BEV feature duplicated from the teacher model. The details of the loss computation are illustrated in the dashed box. The \oplus in the figure denotes the summation of features.

C. Fusion-Based Knowledge Distillation Module

As shown in Fig. 2, our teacher and student model share the same network structure. Because both the fusion and distillation processes are under BEV, the image branching of the student model is optional. In this section, we use $(F_{s,0}, F_{s,1}, F_{s,2}, F_{s,3})$ to represent the four BEV features of the student model, and $(F_{t,0}, F_{t,1}, F_{t,2}, F_{t,3})$ to represent the corresponding features of the teacher model (i.e., multi-scale bev features on a light blue background in Figs. 2 and 3).

1) *Cross-Model Residual Distillation*: Camera and LiDAR fusion features contain rich semantic and geometric information. Compared to them, radar points are much sparser. The semantic information of radar is mainly derived from velocity measurements. Observing this gap, we believe the standard approach of directly forcing radar features to mimic multi-modal features may not work well [24], so we design a Cross-Model Residual Distillation module. Specifically, we use (3) to project student features F_s onto a latent space F'_s with the same dimensions. We address the discrepancies in features resulting from using different feature extractors by reducing the cosine similarity, calculated using (4), between the student and teacher features. Subsequently, we reintegrate F'_s with the original F_s to provide additional information. This approach preserves the intrinsic logic of the radar features and enhances the training process

of the radar backbone, in contrast to direct feature mimicry methods. Moreover, radar possesses a distinct advantage over cameras and LiDAR in terms of weather resilience and observation range. We aim for the student model to learn from the teacher's strengths while maintaining its unique characteristics rather than simply copying the teacher. Based on the above, we utilize ARF to dynamically calculate the weights for integrating F'_s with F_s . The procedure for feature transfer is outlined as follows:

$$F'_s = MLP(F_s), \quad (2)$$

$$F_s^* = F_s + F'_s \times ARF(F'_s). \quad (3)$$

Assuming that $f_{(u,v)}$ is a feature indexed as (u, v) on the feature map F , the CMRD loss \mathcal{L}_{cmrd} is formed as follows:

$$\mathcal{L}_{cmrd} = 1 - \frac{1}{H \cdot W} \sum_u \sum_v M_{u,v} \frac{f'_{s,(u,v)} f_{t,(u,v)}}{\|f'_{s,(u,v)}\|_2 \|f_{t,(u,v)}\|_2} \quad (4)$$

where $M_{u,v} = 1$ if there are labels on that pillar that are non-empty and non-noise, and $M_{u,v} = 0$ otherwise. In other words, we constrain the feature similarity only on occupied locations to minimize computational consumption. As shown in Fig. 3, we computed \mathcal{L}_{cmrd} for $F_{s,i}$ and $F_{t,i}$ for $i = 0, 1, 2$, where the green dashed box shows the details of the computed loss.

2) *BEV Relation Distillation*: This section introduces a mechanism designed to uphold the consistency of scene-level geometric relationships. To achieve this, we employ a cosine similarity-based affinity matrix to compare teacher feature maps F_t and student feature maps F_s . Initially, the tensors F_s and F_t are defined in the space $\mathbb{R}^{C \times H \times W}$. We then transform these tensors into matrices with dimensions $C \times (H \cdot W)$. The affinity matrix is computed using the following calculation:

$$A_{u,v} = \frac{f_u^\top f_v}{\|f_u\|_2 \|f_v\|_2}, (u, v \in \{1, 2, \dots, K = H \cdot W\}), \quad (5)$$

where $A_{u,v}$ denotes the cosine similarity at each element (u, v) of the affinity matrix, f_u denotes the u -th feature in the feature map F . To assess the scene-level information gap between the student and teacher model, we compute the L1 norm between

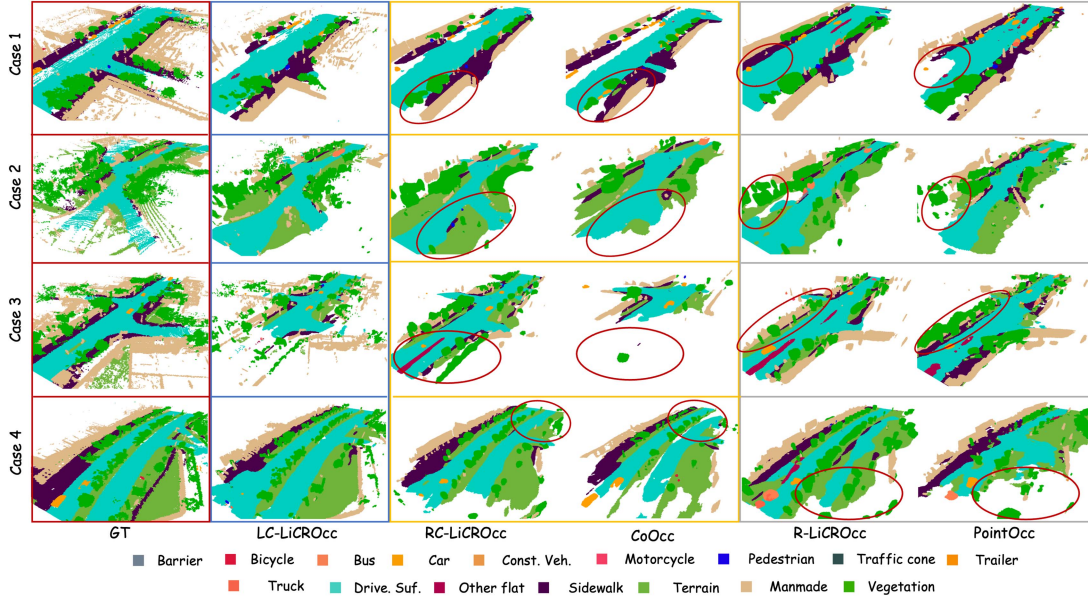


Fig. 4. Visual comparison of our LiCROcc with baseline methods on OpenOccupancy benchmark. Our methods offer a more comprehensive representation of the scene and more accurate segmentation boundaries compared to CoOcc and PointOcc. Surprisingly, the R-LiCROcc can segment feasible regions and obstacles even with only two thousand radar points as input.

their affinity matrices. The BRD loss is then defined as follows:

$$\mathcal{L}_{\text{brd}} = \frac{1}{K \cdot K} \sum_{u=1}^K \sum_{v=1}^K \|A_{u,v}^S - A_{u,v}^T\|_1, \quad (6)$$

where A^S, A^T denote the affinity matrix of the student and teacher network bev feature maps, respectively. As shown in Fig. 3, we computed \mathcal{L}_{brd} for $F_{s,i}$ and $F_{t,i}$ for $i = 1, 2, 3$, where the gray dashed box shows the details of the computed loss. To alleviate the computation burden, we resize all the BEV features of different scales to a smaller resolution and then compute \mathcal{L}_{brd} . It is worth noting that BRD has only been used to distill the radar-camera fusion for the student model.

3) *Predictive Distribution Distillation*: KL divergence can be used to measure the difference between two distributions. Unlike MonoOcc [32] and MonoScene [23], this letter aims to supervise the distillation of KL divergence across multiple modal combinations. Specifically, we compute the KL divergence upon probabilities $\tilde{Y} = \text{softmax}(\mathbf{Y})$, where \mathbf{Y} is introduced in Section III-B, predicted by teacher and student models. This measure captures the distribution discrepancy and is integrated into the distillation objective. By minimizing the KL divergence, the student model is encouraged to align its predictions closely with those of the teacher, thereby enhancing its predictive capabilities. LC-Fusion networks provide denser information compared to radar-camera or radar, offering cues that help the student network mitigate sparsity and improve its performance (as shown in Fig. 4). The PDD loss can be computed as follows:

$$\mathcal{L}_{\text{pdd}} = \text{KL}(\tilde{Y}_S || \tilde{Y}_T), \quad (7)$$

where \tilde{Y}_S is the predictive probability distribution of the student model and \tilde{Y}_T , the learning target, is the predictive probability distribution of the teacher model.

D. Overall Loss Functions

We adopt a multi-task training strategy during the training phase to effectively guide the various components. For 3D SSC, we utilize the $\mathcal{L}_{\text{loss}}$ loss. To enhance the distillation process, we combine three distillation components: $\mathcal{L}_{\text{cmrd}}$, \mathcal{L}_{brd} and \mathcal{L}_{pdd} . Additionally, we retain the semantic (\mathcal{L}_s) and occupancy (\mathcal{L}_c) losses from SSC-RS [15] to supervise feature point cloud extraction. The overall loss function is represented as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{loss}} + \lambda_2 \mathcal{L}_{\text{cmrd}} + \lambda_3 \mathcal{L}_{\text{brd}} + \lambda_4 \mathcal{L}_{\text{pdd}} + \lambda_5 (\mathcal{L}_c + \mathcal{L}_s) \quad (8)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are hyper-parameters.

IV. EXPERIMENTS

In this section, we elaborate on the evaluation datasets and metrics (Section IV-A), implementation details (Section IV-B), and performance comparisons with state-of-the-art (SOTA) methods (Section IV-C). Additionally, we conduct ablation studies to demonstrate the effectiveness of the proposed fusion module (Section IV-D1) and distillation module (Section IV-D2). Finally, we provide experiments to ablate the impact of the observation distance (Section IV-D3) and the unique benefits of radar for SSC (Section IV-D4).

A. Dataset and Metrics

Datasets: We evaluated our method on the nuScenes-Occupancy [3] benchmark, which consists of 700 scenes (70% sunny daytime) for training and 150 scenes (74% sunny daytime) for validation, with 34,000 keyframes on a 3D volume of size $40 \times 512 \times 512$. The dataset covers an area of -51.2 to 51.2 meters in the xy plane and -5 to 3 meters in the z -axis. It provides the voxel annotations with 17 categories and a $0.2 \text{ m} \times 0.2 \text{ m} \times 0.2 \text{ m}$ resolution.

TABLE II
MODEL EFFICIENCY COMPARISON

Method	GFLOPS↓	Memory(GB)↓	Params(M)↓	FPS↑
Co-Occ	739.1	8.2	176.3	0.6
RC-LiCROcc	261.5	5.7	70.7	7.3
PointOcc	203.3	39.8	29.3	7.5
R-LiCROcc	173.6	4.9	24.2	9.3

Metrics: Following [37], we employ Intersection-over-Union (IoU) for scene completion (excluding semantics), and the mean Intersection-over-Union (mIoU) for semantic scene completion (no “noise” class), as the validation metrics.

B. Implementation Details

For LiDAR inputs, we concatenate 10 LiDAR sweeps as a keyframe, with a visual range of $[R_x, R_y, R_z] = [102.4 \text{ m}, 102.4 \text{ m}, 8 \text{ m}]$, similar to [3], [5]. We leverage a pre-trained ResNet50 [38] on ImageNet [39] as the image backbone to process the camera images with the input resolution of 256×704 . During training, we project the point cloud onto the camera view to monitor the depth of the LSS [40] projection module, which serves as the transformer from PV to BEV. For the radar input, we adopted the preprocessing procedure in CRN [6], using radar scans stitched together from a total of 8 sweeps from 5 radar sensors of the car. For data augmentation, we randomly apply flips and cropping to the images. The point clouds are augmented by the random flipping on the x -axis and y -axis. We employ the AdamW [41] optimizer with a weight decay of 0.01 and an initial learning rate of $2e-4$. We use the cosine learning rate scheduler with linear warming up in the first 500 iterations. All experiments are conducted on 8 NVIDIA A100 GPUs with a total batch size 32 for 24 epochs.

C. Quantitative Results

Table I shows the comparison results of our method against the SOTA methods on the nuScenes-Occupancy [3] benchmark. Compared with all previous methods, our LiCROcc achieves the best performance under the same configuration. For example, our LiDAR-camera fusion model LC-Fusion shows significant improvements, with a 23% increase in mIoU and an 18% increase in IoU compared to the baseline (M-CONet [3]). Meanwhile, LC-Fusion achieves improvements of 3.3% and 2.3% in terms of mIoU and IoU scores over PointOcc [5], which emphasizes the effectiveness of our proposed multi-modal BEV fusion.

To comprehensively evaluate the effectiveness of our proposed method, we have modified several existing LiDAR-based and multi-modal methods (CONet [3], SSC-RS [15], PointOcc [5], and CoOCC [4]), to accommodate radar input, serving as our comparisons (efficiency comparison is in Table II). Specifically, we replace the LiDAR input $\mathbf{P}_{lidar} \in \mathbb{R}^{N \times 4}$ for these models with the Radar $\mathbf{P}_{radar} \in \mathbb{R}^{N \times 7}$ input and change the number of input channels of the voxel encoder in the model and leave the rest unchanged. As shown in the second part of Table I, our R-LiCROcc outperforms the second-best one (PointOcc) and the baseline (R-SSC-RS) by 13.3% on the mIoU scores, demonstrating the effectiveness of our proposed fusion-based KD. We find the IoU score of our R-LiCROcc is slightly lower than PointOcc. We explain that this can be

TABLE III
ABLATION ON FUSION STAGES

Input	stages	IoU↑	mIoU↑	Input	stages	IoU↑	mIoU↑
L	0	34.4	22.1	R	0	20.8	7.5
C&L	1	34.0	23.4	C&R	1	24.7	15.0
	2	34.2	23.7		2	24.9	15.4
	3	34.9	24.7		3	25.2	15.6

TABLE IV
ABLATION ON DISTILLATION MODULES

Model	CMRD	BRD	PDD	IoU↑	mIoU↑
R-LiCROcc	✓			20.71±0.10	7.50±0.06
	✓		✓	21.66±0.04	7.87±0.02
RC-LiCROcc				24.87±0.33	15.31±0.25
	✓			25.25±0.04	16.03±0.02
	✓	✓		25.55±0.07	16.22±0.02
	✓	✓	✓	25.97±0.08	17.12±0.01
RC-CONet [3] (8 epochs)				19.7	13.8
	✓			20.9	14.4
		✓		20.7	14.5
			✓	20.1	13.9
Co-Occ [4] (8 epochs)				23.3	15.8
	✓			24.1	15.6
		✓		23.7	15.9
			✓	24.3	15.9

attributed to PointOcc projects features on three planes and uses a larger model, which may be more beneficial for occupancy prediction. For radar-camera fusion, we take CONet and CoOCC as our baselines. Voxel features are employed to represent the scene, and the fusion process of visual and point cloud data is also conducted through 3D operations. In contrast, our approach involves compressing the scene into the BEV and performing the fusion process within the BEV, resulting in improved speed (as depicted in Table II). The results are presented in the third part of Table I and show that our radar-camera fusion version RC-Fusion has achieved comparable performance to these baselines. The proposed fusion-based KD further boosts the performance by 1.5 and 0.8 in terms of mIoU and IoU. We also provide the visualization in Fig. 4 illustrates that our RC-LiCROcc and R-LiCROcc enable complete scene completion and precise object segmentation.

D. Ablation Studies

We conduct a series of experiments to validate the proposed module and the potential of radar as a sensor for SSC tasks. All experiments are conducted under the same training configuration and evaluated according to nuScene-Occupancy [3] validation dataset.

1) *Effect of Fusion Module:* We investigate the impact of different fusion stages in the multi-modal BEV fusion network presented in III-B. Specifically, we fuse camera and point cloud features with dimensions $(C_i \times H_i \times W_i)$ equal to $(64 \times 256 \times 256)$, $(128 \times 128 \times 128)$, and $(256 \times 64 \times 64)$, respectively. The corresponding results are shown in Table III. “Stages=0” means only using the point cloud as input, which serves as the point-based baseline. From Table III, we can see that the multi-stage fusion strategy effectively improves the accuracy of SSC.

TABLE V
PERFORMANCE BREAKDOWN BY RANGE EVALUATED ON THE NUSCENES VAL SPLIT

Method	Modality	[0m, 20m]	IoU↑ [20m, 30m]	[30m, 50m]	[0m, 20m]	mIoU↑ [20m, 30m]	[30m, 50m]
LC-Fusion (Teacher)	L+C	49.0	23.69	12.01	34.62	14.85	4.98
R-SSC-RS	R	28.12	7.07	0.91	9.38	2.25	0.39
R-LiCROcc (Student)	R	27.87(-0.25)	9.56(+2.49)	0.81(-0.1)	10.44(+1.06)	3.26(+1.01)	0.81(+0.42)
RC-Fusion	C+R	35.79	10.56	1.75	21.75	6.39	1.02
RC-LiCROcc (Student)	C+R	36.52(+0.73)	11.67(+1.11)	2.27(+0.52)	23.66(+1.91)	7.27(0.88)	1.22(+0.2)

TABLE VI
PERFORMANCE BREAKDOWN BY WEATHER AND LIGHTING EVALUATED ON THE NUSCENES VAL SPLIT.

Method	Modality	IoU↑				mIoU↑			
LC-Fusion (Teacher)	L+C	35.22	34.28	33.52	34.16	24.91	24.08	16.43	12.69
SSC-RS	L	35.35	32.27	33.16	32.37	21.23	19.80	13.02	10.90
C-CONet	C	23.20	19.90	9.80	9.60	14.10	12.70	4.80	3.64
R-SSC-RS	R	21.25	19.27	18.90	17.01	7.55	7.15	5.35	4.30
R-LiCROcc (Student)	R	21.86(+0.61)	20.10(+0.83)	19.30(+0.4)	17.27(+0.26)	8.59(+1.04)	7.10(-0.05)	5.69(+0.34)	4.74(+0.44)
RC-Fusion	R+C	25.88	24.49	20.47	20.10	15.85	15.30	9.39	7.10
RC-LiCROcc (Student)	R+C	26.72(+0.84)	25.12(+0.63)	21.23(+0.76)	20.67(+0.57)	17.48(+1.63)	16.69(+1.39)	10.10(+0.71)	8.02(+0.92)

2) *Effect of Distillation Module*: In this section, we delve into the individual contributions of different distillation components of our proposed fusion-based KD. Detailed results are illustrated in Table IV. Each configuration starts with a baseline model with any distillation module, to which we sequentially add our distillation modules to evaluate their efficacy. We also evaluated the contributions of each of the three distillation modules on RC-CONet [3] and Co-Occ [4], which we trained for 8 epochs each, observing improvements across all metrics, particularly a more pronounced increase in the IoU. Specifically, we directly compute the CMRD loss in 3D fusion features, and compute BRD loss on BEV features compressed with maxpool. Results in both parts of Table IV show that CMRD, BRD, and PDD components significantly enhance the performance. Among them, PDD provides the greatest improvement in the accuracy of the model's semantic classification, registering a 8.6% mIoU improvement for R-LiCROcc and a 5.5% mIoU improvement for RC-LiCROcc, underscoring its crucial role in cross-modal KD. However, the limited density of the radar point cloud results in minimal efficacy for PDD in predicting radar occupancy.

3) *Visual Field Benefits From KD*: Radar's inherent ability to penetrate objects and bypass foreground obstacles enables it to provide a wider field of view than LiDAR and camera sensors. However, the sparsity of radar point clouds increases with distance, which is particularly unfavorable for SSC, as shown in rows 2 and 3 of Table V.

In order to further analyze the improvements brought by our KD modules, we conducted a statistical analysis to evaluate the system's effectiveness for semantic scene completion across various distance ranges, as detailed in Table V. We measure the IoUs and mIoUs of the teacher model, the student model, and the R-LiCROcc at [0 m, 20 m], [20 m, 30 m], and [30 m, 50 m] for SSC, respectively. Table V reveals that KD significantly improves the performance of semantic categorization of student models, particularly in the short-range area. However, since the radar near point cloud comprises only 12.2% of the total points, improving the radar's capability for occupancy prediction (i.e., IoU metrics) is challenging. Interestingly, we found that when performing KD from LiDAR-camera fusion to radar-based models, the improvement in mIoU scores for the long-range

area is much smaller than that for short- and medium-range areas. This observation suggests that the LiDAR-camera fusion loses its advantage as distance increases due to its shorter visual range. It is worth noting that both the teacher and student models exhibit severe performance degradation in the long-range area, especially in the mIoU score. For example, the teacher model outperforms RC-LiCROcc by 10.96 mIoU within 20 m. However, this advantage sharply drops to 3.76 points (almost 65% decrease) in the [30 m, 50 m] range.

4) *Weather Robustness From Radar*: This study evaluates the performance of the radar-based methods in various weather conditions. Results detailed in Table VI reveal that models' performances fluctuate with changing weather scenarios (sunny daytime, rainy day, nighttime, and rainy night).

First, as shown in Table VI, the weather properties of the three sensor types reveal varying degrees of robustness. The mIoU for radar decreases by only 3.25 from a clear day to a rainy night, while for LiDAR and camera, it decreases by 10.33 and 10.46 points, respectively. This indicates that radar is the most resilient to adverse weather and lighting conditions. In particular, during clear daylight hours, the teacher model achieves 16.32 higher than R-LiCROcc and 7.43 higher than RC-LiCROcc in terms of mIoU scores. However, in rainy night conditions, this advantage narrows to 7.95 and 5.26, respectively, with dominant performance decreasing by 51.3% and 29.2%. Additionally, the rain in the nuScenes dataset is not particularly heavy, resulting in a less significant impact on the LiDAR point cloud than anticipated. Examining radar performance under a broader range of weather conditions is a focus of our future work.

Under sunny daytime conditions, the distillation effect yields the highest performance. The R-LiCROcc model demonstrates a 2.8% improvement in IoU and a 13.8% improvement in mIoU compared to the student model. Similarly, the RC-LiCROcc model achieves a 3.2% increase in IoU and a 10.3% increase in mIoU. This enhancement is attributed to the optimal performance of the teacher model under sunny conditions. Conversely, during rainy days and nights, the visibility of both the LiDAR and camera is compromised, leading to less pronounced enhancements for the student model. In fact, the performance of the R-LiCROcc model is slightly diminished in rainy weather.

V. CONCLUSIONS AND FUTURE WORK

In this letter, we investigate the utilization of radar in SSC tasks. We initially developed a fusion network that integrates point clouds and images, complemented by three distillation modules. By leveraging the strengths of radar while augmenting its performance on the SSC task, our approach achieves superior results across diverse settings. We plan to examine the SSC performance of additional radar types to substantiate our findings.

REFERENCES

- [1] Z. Ming, J. S. Berrio, M. Shan, and S. Worrall, "OccFusion: Multi-Sensor Fusion Framework for 3D Semantic Occupancy Prediction," *IEEE Trans. Intell. Veh.*, 2024.
- [2] H. Li et al., "Riders: Radar-infrared depth estimation for robust sensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 18764–18778, 2024, doi: [10.1109/TITS.2024.3432996](https://doi.org/10.1109/TITS.2024.3432996).
- [3] X. Wang et al., "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17850–17859.
- [4] J. Pan, Z. Wang, and L. Wang, "Co-Occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3D semantic occupancy prediction," *IEEE Robot. Automat. Lett.*, vol. 9, no. 6, pp. 5687–5694, Jun. 2024.
- [5] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "PointOCC: Cylindrical TRI-perspective view for point-based 3D semantic occupancy prediction," 2023, *arXiv:2308.16896*.
- [6] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "CRN: Camera radar net for accurate, robust, efficient 3D perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17615–17626.
- [7] Y. Long, A. Kumar, D. Morris, X. Liu, M. Castro, and P. Chakravarty, "Radiant: Radar-image association network for 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1808–1816.
- [8] F. Ding, X. Wen, Y. Zhu, Y. Li, and C. X. Lu, "RadarOCC: Robust 3D occupancy prediction with 4D imaging radar," *Adv. Neural Inf. Process. Syst.*, 2024.
- [9] G. Bang, K. Choi, J. Kim, D. Kum, and J. W. Choi, "Radardistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15491–15500.
- [10] M. Klingner et al., "X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13343–13353.
- [11] L. Zheng et al., "RcFusion: Fusing 4-D radar and camera with bird's-eye view features for 3-D object detection," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 8503814.
- [12] 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 Trans. Intell. Veh.*, vol. 9, no. 1, pp. 79–92, Jan. 2024.
- [13] L. Wang et al., "Interfusion: Interaction-based 4D radar and LiDAR fusion for 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 12247–12253.
- [14] R. Xu and Z. Xiang, "Rlnet: Adaptive fusion of 4D radar and LiDAR for 3D object detection," in *Proc. ROAM ECCV*, 2024.
- [15] J. Mei, Y. Yang, M. Wang, T. Huang, X. Yang, and Y. Liu, "SSC-RS: Elevate LiDAR semantic scene completion with representation separation and BEV fusion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 1–8.
- [16] Z. Xia et al., "SCPNet: Semantic scene completion on point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17642–17651.
- [17] A.-Q. Cao, A. Dai, and R. de Charette, "Pasco: Urban 3D panoptic scene completion with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14554–14564.
- [18] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3CNet: A sparse semantic scene completion network for LiDAR point clouds," in *Proc. Conf. Robot Learn.*, 2021, pp. 2148–2161.
- [19] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrilu, "Semantic scene completion using local deep implicit functions on LiDAR data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7205–7218, Oct. 2022.
- [20] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9223–9232.
- [21] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9433–9443.
- [22] Z. Yu et al., "FlashOcc: Fast and memory-efficient occupancy prediction via channel-to-height plugin," 2023, *arXiv:2311.12058*.
- [23] A.-Q. Cao and R. De Charette, "MonoScene: Monocular 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3991–4001.
- [24] L. Zhao, J. Song, and K. A. Skinner, "CRKD: Enhanced camera-radar object detection with cross-modality knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15470–15480.
- [25] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, "RCM-Fusion: Radar-camera multi-level fusion for 3D object detection," in *2024 IEEE Int. Conf. Robot. Automat.*, 2024, pp. 18236–18242.
- [26] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, "Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1523–1535, Feb. 2023.
- [27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [28] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, "Unidistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5116–5125.
- [29] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Bevdistill: Cross-modal BEV distillation for multi-view 3D object detection," 2022, *arXiv:2211.09386*.
- [30] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for LiDAR semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8479–8488.
- [31] X. Yan et al., "2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 677–695.
- [32] Y. Zheng et al., "MonoOcc: Digging into monocular semantic occupancy prediction," in *Proc. 2024 IEEE Int. Conf. Robot. Automat. (ICRA)*, 2024, pp. 18398–18405, doi: [10.1109/ICRA57147.2024.10611261](https://doi.org/10.1109/ICRA57147.2024.10611261).
- [33] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3D sketch-aware semantic scene completion via semi-supervised structure prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4193–4202.
- [34] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3351–3359.
- [35] L. Roldão, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3D semantic completion," in *Proc. 2020 Int. Conf. 3D Vis. (3 DV)*, 2020, pp. 111–119.
- [36] X. Yan et al., "Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3101–3109.
- [37] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1746–1754.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [40] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Comput. Vis.—ECCV 2020: 16th Eur. Conf., Glasgow, U.K., Aug. 23–28, 2020, Proc., Part XIV 16*, Springer, 2020, pp. 194–210.
- [41] I. Loshchilov et al., "Fixing weight decay regularization in adam," 2017, *arXiv:1711.05101*.