

# RIOcc: Efficient Cross-Modal Fusion Transformer with Collaborative Feature Refinement for 3D Semantic Occupancy Prediction

Baojie Fan<sup>1\*</sup> Xiaotian Li<sup>1</sup> Yuhan Zhou<sup>1</sup> Yuyu Jiang<sup>1</sup> Jiandong Tian<sup>2</sup> Huijie Fan<sup>2</sup>

<sup>1</sup>Nanjing University of Posts and Telecommunications

<sup>2</sup>Shenyang Institute of Automation, Chinese Academy of Sciences

{jobfbj, xiaotianli981, yuhanzhou521, jyy01731}@gmail.com {tianjd, fanhuijie}@sia.cn

## Abstract

The multi-modal 3D semantic occupancy task provides a comprehensive understanding of the scene and has received considerable attention in the field of autonomous driving. However, existing methods mainly focus on processing large-scale voxels, which bring high computational costs and degrade details. Additionally, they struggle to accurately capture occluded targets and distant information. In this paper, we propose a novel LiDAR-Camera 3D semantic occupancy prediction framework called RIOcc, with collaborative feature refinement and multi-scale cross-modal fusion transformer. Specifically, RIOcc encodes multi-modal data into a unified Bird’s Eye View (BEV) space, which reduces computational complexity and enhances the efficiency of feature alignment. Then, multi-scale feature processing substantially expands the receptive fields. Meanwhile, in the LiDAR branch, we design the Dual-branch Pooling (DBP) to adaptively enhance geometric features across both the Channel and Grid dimensions. In the camera branch, the Wavelet and Semantic Encoders are developed to extract high-level semantic features with abundant edge and structural information. Finally, to facilitate effective cross-modal complementarity, we develop the Deformable Dual-Attention (DDA) module. Extensive experiments demonstrate that RIOcc achieves state-of-the-art performance, with 54.2 mIoU and 25.9 mIoU on the Occ3D-nuScenes and nuScenes-Occupancy datasets, respectively.

## 1. Introduction

3D semantic occupancy prediction is a complex and vital task that aims to jointly estimate the geometric structure and semantic categories of voxels within a scene. Compared to traditional tasks [13, 25, 37, 42] like 3D object detection, 3D semantic occupancy prediction focuses more on understanding the overall scene, which includes not only objects

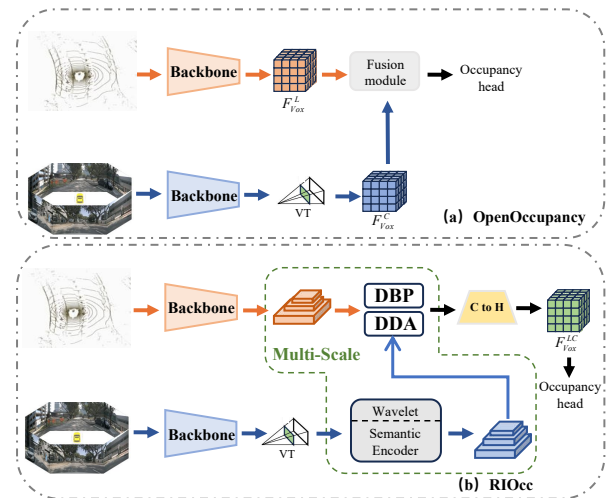


Figure 1. Comparison between OpenOccupancy and the proposed RIOcc. Instead of processing voxel features like OpenOccupancy, we choose BEV features to achieve higher computational efficiency. Additionally, we extracted refined multi-modal BEV features and fed them into the efficient fusion module (DDA). The fused BEV features are then transformed into 3D features for occupancy prediction. “VT” is view transform.

but also details in the background and environment. Consequently, it provides richer and more reliable environmental perception, which is essential for applications such as high-precision map construction and path planning in autonomous driving [5, 50].

In contrast to traditional methods that rely solely on a single modality [19, 22, 23, 51], multi-modal fusion [1, 6, 15, 41, 59] provides higher accuracy and robustness for 3D perception. In various 3D perception tasks, effectively combining data from cameras and LiDAR presents a crucial challenge for achieving high-precision predictions. Cameras provide rich semantic and texture information, while LiDAR captures accurate geometric and depth information. The complementarity of the two has driven research on fusion strategies to improve the performance of 3D semantic

\*Corresponding Author.

occupancy prediction. Nevertheless, due to the heterogeneity of camera and LiDAR data, along with the complexity of feature integration, interactions between different modalities often lead to inconsistencies in occupancy prediction. It can degrade semantic information and introduce significant uncertainty into the predictions.

However, the task of semantic occupancy prediction [2, 9, 10, 12, 39, 40, 49] also faces significant computational challenges, especially when it involves real-time processing of large-scale voxel data, which demands high computational resources. To manage this computational burden, previous single-modal approaches [7, 17, 18, 21, 24, 55, 57] processed data from a Bird’s Eye View (BEV) perspective and achieved considerable results. While multi-modal occupancy methods can significantly improve performance compared to previous approaches, they also increase computational complexity. To find a better balance between computational efficiency and perception accuracy in multi-modal methods, reasonable and effective design is essential. Additionally, data imbalance in complex environments hinders models from learning rare but important spatial features, which is critical issue that needs to be addressed.

To address the aforementioned issues, we propose RIOcc, a novel multi-modal 3D semantic occupancy prediction method. The comparison with the pioneering work OpenOccupancy is shown in Figure 1. RIOcc avoids reliance on 3D convolutions, enabling the extraction of finer features from the BEV perspective and enhancing information complementarity between different modalities. Specifically, we first extract features from LiDAR and camera data and project them into a unified BEV space. To reduce the impact of redundant features on overall performance while preserving rich geometric depth information, we design Dual-branch Pooling (DBP). In addition, the Wavelet Encoder can effectively capture important edge and structural features in images. It compresses data in the channel dimension, which helps reduce computational complexity. The Semantic Encoder is employed to extract high-level features with rich semantic information from the input data. These features are then combined with geometric features in multi-modal fusion, enhancing the comprehensive understanding of the scene.

In the feature fusion stage, RIOcc employs Deformable Dual-Attention (DDA) to facilitate interaction and dynamic adjustment of features from LiDAR and camera. This approach significantly bridges the differences between modalities and enhances information exchange and prediction consistency. Through the DDA module, our model can leverage data from multiple modalities to achieve a comprehensive understanding of both the geometric and semantic aspects of the scene. Finally, the integrated features are fed into the occupancy prediction module. Similar to FlashOcc, we convert the fused BEV features from the channel dimen-

sion to the height dimension. This operation provides semantic labels for each voxel and generates a 3D scene representation rich in semantic information.

Our contributions are summarized as follows:

- We propose a novel multi-modal 3D semantic occupancy prediction framework, RIOcc. The LiDAR and camera branches respectively extract refined structural information and semantic features, with balanced computational load and performance.
- The proposed Deformable Dual-Attention facilitates multi-modal feature fusion, bridging the gaps caused by modality differences, which enhances semantic consistency and reduces modality heterogeneity.
- We conduct extensive experiments on the nuScenes-Occupancy and nuScenes-Occ3d dataset, which demonstrate the effectiveness of our method and achieve state-of-the-art performance.

## 2. Related Work

**BEV-based 3D Perception.** In recent years, BEV perception has become crucial for autonomous driving due to its global view and efficiency. With declining camera costs, vision-based 3D object detection has rapidly advanced. A key work is LSS [34], which uses geometric methods to create depth maps. This approach was first integrated into a full detection pipeline in BEVDet [13]. Later, BEVDepth [22] added explicit depth supervision, and Far3D [16] introduced sparse queries and adaptive 3D query generation. However, single-modal methods still struggle with accurate depth and spatial understanding, leading to a growing focus on multi-modal BEV perception. BEVFusion [27] was a pioneering multi-modal approach, projecting different features into BEV for unified fusion. MaskBEV [59] further added task-agnostic masks to the BEV space. While effective, these methods still face challenges in detecting non-target objects.

**Camera-based 3D Semantic Occupancy Prediction.** To overcome the intrinsic limitations of traditional object detection approaches, 3D semantic occupancy prediction has gained prominence. MonoScene [5] leverages optically inspired 2D-3D feature projection to infer dense geometric and semantic information. TPVFormer [14] employs a three-view representation combined with Transformers to project 2D image features into 3D space for prediction. Recently, OccupancyM3D [33] directly learns occupancy states in 3D space, and integrates occupancy information into the feature representation of monocular 3D detection. PanoOcc [48] employs voxel queries and a coarse-to-fine approach, effectively processing semantic information for all objects in complex scenes. However, due to inaccuracies in depth estimation and insufficient geometric information in pure visual occupancy, incorporating multi-modal perception presents a more reliable solution.

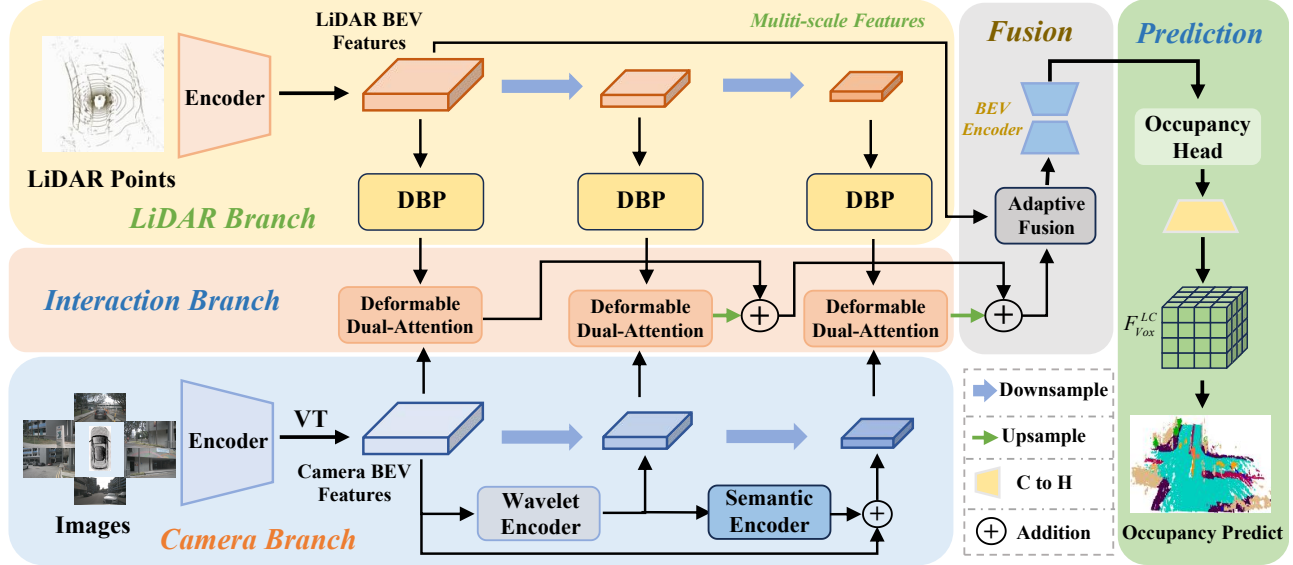


Figure 2. **The overall framework of RIOcc.** This framework includes three main branches: LiDAR, Camera, and Interaction Branch. The LiDAR Branch processes LiDAR points through an encoder and DBP modules to produce multi-scale BEV features. The Camera Branch encodes multi-view images, refining the output with Wavelet and Semantic Encoders to capture spatial and semantic details. In the Interaction Branch, Deformable Dual-Attention integrates features from both branches. These multi-scale features are then adaptively fused and passed to an Occupancy Head for the final occupancy prediction.

**Multi-modal 3D Semantic Occupancy Prediction.** To further improve the precision of 3D semantic occupancy prediction, researchers have increasingly focused on exploring multi-modal perception approaches. OpenOccupancy [47] introduces the first benchmark model for LiDAR-camera semantic occupancy. Building upon this foundation, subsequent work OccGen [46] proposes a “noise-to-occupancy” generative paradigm, which iteratively removes random Gaussian noise to generate and refine 3D semantic occupancy maps. The Co-Occ [31] method presents a multi-modal 3D semantic occupancy prediction framework that combines explicit features fusion with volume rendering regularization. OccFusion [30] improves the utilization of information from different sensors through a dynamic fusion module. OccMamba [20] introduces the Mamba architecture and a 3D-to-1D reordering strategy to enhance the efficiency of processing large-scale 3D voxels. Despite significant advancements in multi-modal perception, challenges like limited interaction, high computational load, and coarse feature representations persist, requiring further optimization.

### 3. Method

#### 3.1. Overall Architecture

The overall architecture of RIOcc is illustrated in Figure 2. The framework takes images and LiDAR point clouds as inputs, extracting consistent BEV features for subsequent fusion (Sec 3.2). The designed Dual-branch Pooling (DBP)

removes redundant features while retaining geometric depth information from the LiDAR branch (Sec 3.3). Simultaneously, to extract more effective camera BEV features, a Wavelet Encoder is employed to suppress noise and enhance hierarchical information. The Semantic Encoder is used to enrich semantic information, improving the understanding of the scene (Sec 3.4). Then, we design the Deformable Dual-Attention (DDA) to strengthen the interaction of BEV features at different scales between modalities (Sec 3.5). Finally, the fused multi-scale features are input into the occupancy prediction module (Sec 3.6).

#### 3.2. Features Extraction

During the feature extraction stage, we design LiDAR and camera branches to encode multi-modal input, following the BEVFusion [25] setup. The LiDAR branch voxelizes point clouds and uses a 3D sparse convolutional network to generate BEV features  $F_L^{BEV}$ . The camera branch extracts multi-scale image features using ResNet50 and maps them to the BEV space using a view transformation, resulting in feature  $F_C^{BEV}$ . Finally, the features from the two branches are fused in the BEV perspective, providing basic feature representations for subsequent occupancy prediction.

#### 3.3. Dual-branch Pooling

To enhance the feature representation of the LiDAR branch, we design Dual-branch Pooling (DBP), which primarily focuses on adaptively refining important geometric features and removing redundant features. Firstly, in order to en-

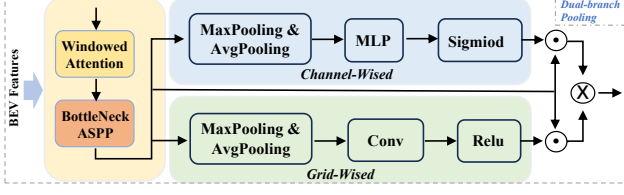


Figure 3. **The schema of Dual-branch Pooling (DBP).** LiDAR feature representation is improved by adaptively highlighting important semantic channels and significant geometric regions.

hance the ability to capture long-range semantics and multi-scale spatial information of global BEV features, the input feature  $F'_L$  passes through the Windowed Attention module followed by the Bottleneck ASPP module. Then, the features are passed through the Channel-wise Attention and Grid-wise Attention modules, optimizing information representation across different dimensions. The detailed structures of DBP are shown in Figure 3.

To adaptively highlight channels that are crucial for semantic information, the channel-wise attention utilizes max pooling and average pooling to aggregate the spatial dimensions of the input BEV features  $F'_L$ , generating two descriptors  $F_{Avg}$  and  $F_{Max}$ . They are processed through a shared MLP to produce a channel attention weight map, which is then activated by a sigmoid function to emphasize the important channel information within the features. The output from the Channel-wise Attention are given by:

$$F_{channel} = \sigma(MLP(F_{Avg}) + MLP(F_{Max})) \quad (1)$$

To focus on regions that are particularly important for the overall geometric description, the spatial attention module also utilizes max pooling and average pooling operations to the input BEV features to compress them along the channel dimension to obtain  $F'_{Avg}$  and  $F'_{Max}$ . Subsequently, these compressed features are processed through convolutional layers and ReLU to generate a spatial attention weight map, which identifies positions that are significant in terms of geometric structure. The features outputted from the Grid-wise Attention can be represented as:

$$F_{grid} = \sigma(ReLU(f^{7 \times 7}(F'_{Avg}; F'_{Max}))) \quad (2)$$

The final output  $F_{DBP}$  can be represented as:

$$F_{DBP} = (F_{channel} \odot F'_L) \otimes (F_{grid} \odot F'_L) \quad (3)$$

where  $\sigma$  denotes the sigmoid activation operation, and  $f^{7 \times 7}$  represents a convolution with a kernel size of  $7 \times 7$ .

### 3.4. Efficiency Camera BEV Features

#### 3.4.1. Wavelet Encoder

Due to differences in viewpoints and hardware characteristics among the different cameras, various noises and inconsistencies are often introduced. To reduce redundancy,

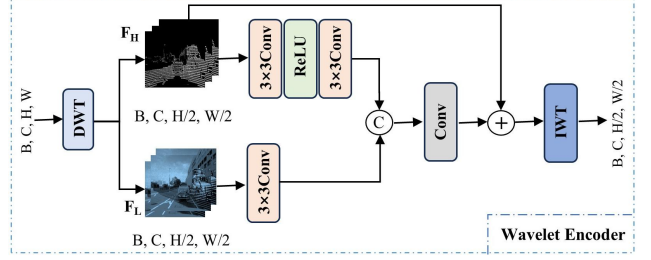


Figure 4. **Detailed structure diagram of the wavelet encoder.** The input BEV features undergo DWT and IWT to obtain richer structure and details.

noise impact, and decrease computational burden, we design the Wavelet Encoder, as shown in Figure 4. This encoder retains the primary structure and details of the images by decomposing features, which enhances the efficiency and accuracy of the fusion stage. Specifically, the Wavelet Encoder first applies a wavelet transform to the input camera BEV features, decomposing them into low-frequency and high-frequency components. The low-frequency components typically contain the main structural information and global features of the images, while the high-frequency components capture details such as edges and textures. This approach allows for a more refined multi-scale decomposition of features, leading to rich semantic and geometric details. We apply Discrete Wavelet Transform (DWT) to the input features  $F_C^{BEV}$  to obtain low-frequency features  $F_{low}$  and high-frequency features  $F_{high}$ :

$$F_C^{BEV} \xrightarrow{DWT} \{F_{low}, F_{high}\} \quad (4)$$

Next, the low-frequency features  $F_{low}$  are retained for subsequent feature fusion, while the high-frequency features  $F_{high}$  undergo further encoding through a series of convolutional layers and nonlinear activation to extract high-contrast information and to eliminate noise. Finally, we concatenate the processed high-frequency features with the low-frequency features and then passed through the Inverse Discrete Wavelet Transform (IWT) to form a multi-scale feature representation  $F_{wavelet}$ . This process can be roughly expressed as:

$$F_{wavelet} \xrightarrow{IWT} Concat(F_{low}, Conv(F_{high})) \oplus F_H \quad (5)$$

#### 3.4.2. Semantic Encoder

To enhance the semantic expressiveness of the BEV features, we propose a lightweight 2D Semantic Encoder for efficiently extracting rich semantic information. This encoder employs a lightweight 2D U-Net [36] architecture, utilizing multi-scale feature downsampling and upsampling to extract and fuse both global and local information. The Semantic Encoder first downsamples the input BEV features to capture global contextual information. Subsequently, it upsamples to restore the spatial dimensions of the

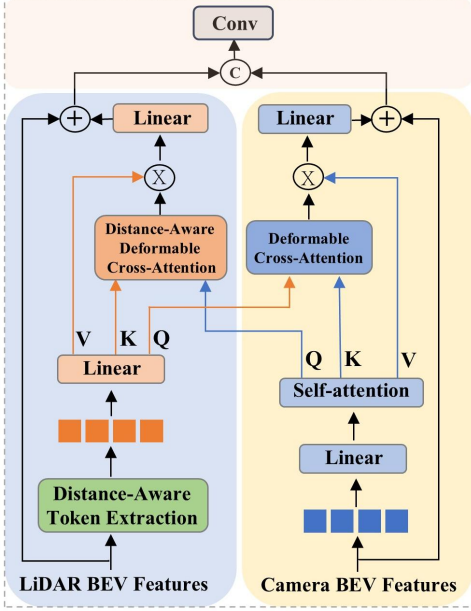


Figure 5. **Overview of Deformable Dual-Attention (DDA)**, which reduces the disparity between LiDAR and Camera BEV features and enhances scene understanding.

features and utilizes residual connections to fuse features of different scales. Additionally, we introduce an Auxiliary Semantic Loss at the output stage to enhance the semantic consistency of the features and improve the model’s understanding of complex scenes.

### 3.5. Deformable Dual-Attention

Due to the inherent sparsity of point clouds, certain regions in the LiDAR BEV representation lack sufficient geometric details, particularly at longer distances. Conversely, although image BEV features provide extensive semantic and texture information, they exhibit limitations in depth estimation accuracy. To enhance distant perception and aggregate local information, we propose a multi-modal fusion transformer that combines the geometric depth information from LiDAR with the semantic information from cameras, which enhance the perception of distant objects for more accurate 3D scene understanding.

As shown in Figure 5, the LiDAR BEV features  $F_{LiDAR} \in \mathbb{R}^{H \times W \times C}$  and Camera BEV features  $F_{Camera} \in \mathbb{R}^{H \times W \times C}$  are used as inputs. For the LiDAR stream, we specially compute the distance  $d_r$  from each LiDAR point to the sensor and encode it using a Gaussian function to generate a distance vector  $e_d$ :

$$e_d = \text{GaussianEncoding}(d_r) \quad (6)$$

Next, we apply a linear transformation using the learned weight vector  $w_{linear}$  and the bias  $b$  obtained through back-propagation to compute a distance-weighted value  $w_{i,j}$ ,

where  $(i, j)$  corresponds to the index of the BEV grid.

$$w_{i,j} = e_d \cdot w_{linear} + b \quad (7)$$

$$F_{LiDAR}^{enhanced}(i, j) = w_{i,j} \cdot F_{LiDAR}(i, j) \quad (8)$$

It allows us to enhance the LiDAR BEV features, compensating for distant features. To enhance the correlation of adjacent spatial features, the features from the  $k$  local regions  $region_k$  within  $F_{LiDAR}^{enhanced}$  are aggregated and to extract a single feature vector. Finally, we extract a set of representative tokens from the enhanced features  $T_{LiDAR}^{(k)}$ . The specific formula is as follows:

$$t_{LiDAR}^{(k)} = \text{Aggregate}(F_{LiDAR}^{enhanced}, region_k) \quad (9)$$

$$T_{LiDAR} = \{t_{LiDAR}^1, t_{LiDAR}^2, \dots, t_{LiDAR}^N\} \quad (10)$$

where *Aggregate* combines multiple features from local regions into a single feature vector. Subsequently, these enhanced tokens are processed through a linear layer to obtain  $Q_l$ ,  $K_l$ , and  $V_l$  for the attention mechanism.

For the Camera stream, the input features are tokenized and undergo self-attention, generating updated  $Q_c$ ,  $K_c$ , and  $V_c$ , which allow for a comprehensive understanding of different regions. Then,  $Q_l$  interacts with  $K_c$ , and  $V_c$  through a deformable cross-modal attention, and vice versa. The deformable cross-attention dynamically captures inter-modal relationships by flexibly selecting feature from the target modality. It effectively focuses on the complementary information between the LiDAR BEV features and the dense camera semantic features. After cross-modal interaction, the LiDAR and Camera BEV features are each combined with their respective linear weights and then concatenated. The DDA module effectively enhances feature interaction between the LiDAR and camera modalities, providing abundant global features for subsequent predictions.

### 3.6. Occupancy Prediction Module

In our framework, the BEV features obtain from the multi-scale fusion stage are input into the occupancy prediction module. These fused features contain rich semantic information and fine-grained geometric structures, which allow for a more comprehensive description of the scene information. Similar to FlashOcc [53], the fused BEV features are processed through the occupancy prediction head before being passed to the Channel-to-Height module. This module rearranges the features from the shape  $C \times W \times H$  to  $C^* \times Z \times W \times H$ , where  $C$ ,  $C^*$ ,  $W$ ,  $H$ ,  $Z$  represent the number of channels, the number of categories, and the dimensions in the three-dimensional space  $(x, y, z)$ . This transformation enables intuitive semantic classification and occupancy prediction of the features in 3D space, significantly enhancing the model’s ability to express occupancy situations in the scene.

Method	Modality	Resolution	Image Backbone	mIoU	IoU																
					others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
LangOcc [3]	C	256×704	R50	11.84	0.00	3.10	90.00	6.30	14.20	0.40	10.80	6.20	9.00	3.80	10.70	43.70	2.23	9.50	26.40	19.60	26.40
FB-Occ [24]	C	256×704	R50	23.12	0.04	37.15	16.81	34.17	38.22	13.41	16.97	19.69	18.94	11.65	21.94	55.94	26.98	29.65	26.92	10.24	14.33
UniVision* [11]	C	256×704	R50	37.50	11.00	44.70	23.10	43.00	50.50	21.60	24.90	26.90	25.70	30.70	35.80	79.80	41.40	49.10	53.80	40.30	34.70
GEOcc* [43]	C	256×704	R50	43.64	<b>14.29</b>	51.27	31.11	46.13	55.09	29.12	30.46	30.99	35.47	35.20	41.82	84.00	47.00	55.52	59.50	50.03	44.82
TEOcc* [26]	C&R	256×704	R50	42.90	10.82	50.33	24.28	48.99	57.32	29.38	24.41	30.14	28.46	36.46	43.01	83.96	43.09	56.00	59.34	54.18	49.16
EFFOcc* [38]	C&L	256×704	R18	49.29	10.57	56.16	21.73	58.68	63.16	31.98	37.71	55.40	36.15	45.87	50.81	81.02	39.07	53.08	57.15	70.41	68.90
OccFormer [58]	C	900×1600	R101	21.93	5.94	30.29	12.32	34.40	19.17	14.44	16.45	17.22	9.27	13.90	26.34	50.99	30.96	34.66	22.73	6.76	6.97
RenderOcc [32]	C	900×1600	R101	26.11	4.84	31.72	10.72	27.67	36.45	13.87	18.20	17.67	17.84	21.19	23.25	63.20	36.42	46.21	44.26	19.58	20.72
TPVFormer [14]	C	900×1600	R101	28.34	6.67	39.20	14.24	41.54	46.98	19.21	22.64	17.87	14.54	30.20	35.51	56.18	33.65	35.69	31.61	19.97	16.12
CTF-Occ [45]	C	928×1600	R101-DCN	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
PanoOcc* [48]	C	640×1600	R101-DCN	42.13	11.67	50.48	29.64	49.44	55.52	23.29	33.26	30.55	30.99	34.43	42.57	83.31	44.23	54.40	56.04	45.94	40.40
OctreeOcc* [28]	C	900×1600	R101-DCN	44.02	11.96	51.70	29.93	53.52	56.77	30.83	33.17	30.65	29.99	37.76	43.87	83.17	44.52	55.45	58.86	49.52	46.33
OccFusion* [30]	C&L	900×1600	R101	46.79	11.65	47.81	32.07	57.27	57.51	31.80	40.11	47.35	33.74	45.81	50.35	78.79	37.17	44.36	53.36	61.18	63.20
OccFusion* [56]	C&L	900×1600	R101	48.74	12.35	51.77	33.01	54.56	57.65	33.99	43.03	48.35	35.54	41.22	48.55	83.00	<b>44.65</b>	57.13	60.01	62.46	61.25
RadOcc* [54]	C&L	512×1408	Swin-B	49.38	10.93	58.23	25.01	57.89	62.85	34.04	33.45	50.07	32.05	48.87	52.11	82.90	42.73	55.27	58.34	68.64	66.01
RIOcc* (Ours)	C&L	256×704	R50	<b>54.21</b>	11.82	<b>59.73</b>	<b>36.98</b>	<b>62.21</b>	<b>68.72</b>	<b>36.45</b>	<b>47.45</b>	<b>58.25</b>	<b>44.20</b>	<b>49.92</b>	<b>54.39</b>	<b>85.10</b>	44.60	<b>59.67</b>	<b>61.77</b>	<b>70.51</b>	<b>69.56</b>

Table 1. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset.** \* means the performance using the camera mask during training. C, L, and R represent camera, LiDAR, and radar, respectively.

Method	Modality	Resolution	Image Backbone	LiDAR Backbone	IoU	mIoU	IoU																
							barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation	
MonoScene [5]	C	900×1600	R101-DCN	-	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	
C-CONet [47]	C	900×1600	R50	-	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	
SparseOcc [44]	C	704×256	R50	-	21.8	14.1	16.1	9.3	15.1	18.6	7.3	9.4	11.2	9.4	7.2	13.0	31.8	21.7	20.7	18.8	6.1	10.6	
LMSCNet [35]	L	-	-	VoxelNet	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 [50]	L	-	-	VoxelNet	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	
L-CONet [47]	L	-	-	VoxelNet	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	
PointOcc [60]	L	-	-	VoxelNet	34.1	23.9	24.9	19.0	20.9	25.7	13.4	<b>25.6</b>	30.6	17.9	16.7	21.2	36.5	<b>25.6</b>	<b>25.7</b>	24.9	24.8	<b>29.0</b>	
M-CONet [47]	C&L	900×1600	R50	VoxelNet	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 [31]	C&L	900×1600	R101	VoxelNet	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	
OccGen [46]	C&L	896×1600	R50	VoxelNet	30.3	22.0	24.9	16.4	22.5	26.1	14.0	20.1	21.6	14.6	17.4	21.9	35.8	24.5	24.7	24.0	20.5	23.5	
OccFusion [56]	C&L	900×1600	R101	VoxelNet	32.4	22.4	25.3	17.0	22.5	25.9	16.5	22.4	24.0	16.1	16.0	22.1	35.6	22.1	24.0	23.9	21.3	24.0	
EFFOcc [38]	C&L	256×704	R18	VoxelNet	30.8	22.9	28.1	16.7	22.1	27.3	13.0	24.8	<b>36.2</b>	<b>22.6</b>	16.8	21.6	29.4	13.9	18.2	20.6	26.5	28.8	
OccMamba [20]	C&L	900×1600	R50	VoxelNet	33.7	25.1	29.6	<b>20.2</b>	25.7	28.5	16.7	25.0	23.2	19.9	20.3	24.5	36.1	25.3	25.1	24.8	<b>27.7</b>	28.9	
RIOcc (Ours)	C&L	256×704	R50	VoxelNet	<b>35.4</b>	<b>25.9</b>	<b>30.2</b>	19.8	<b>25.8</b>	<b>28.7</b>	<b>18.3</b>	24.8	31.8	21.8	<b>20.5</b>	<b>24.9</b>	<b>37.2</b>	24.5	25.5	<b>24.9</b>	27.0	28.8	

Table 2. **3D Occupancy prediction performance on nuScenes-Occupancy validation set.** C represents camera and L represents LiDAR.

### 3.7. Loss

To effectively train the proposed semantic occupancy prediction framework, we combine various types of losses to comprehensively optimize network performance. The cross-entropy loss  $L_{ce}$  and Lovasz-Softmax loss  $L_{ls}$  are used to optimize the overall framework. Affinity loss  $L_{geo}$  and  $L_{sem}$  are applied to optimize scene-wise and class-wise metrics, while  $L_d$  provides feedback for the depth-aware view transform module. Additionally, we introduce an Auxiliary Semantic Loss  $L_{aux}$  to optimize the refined semantic features extracted by the semantic encoder. Therefore, the overall loss function can be expressed as:

$$L_{total} = L_{ce} + L_{ls} + L_{geo} + L_{sem} + L_d + L_{aux} \quad (11)$$

## 4. Experiment

### 4.1. Dataset and Metrics

**Dataset.** Similar to previous works [29, 31, 47, 52], we conducted extensive experiments based on the nuScenes dataset. The dataset is divided into nuScenes-occupancy [4] and Occ3D-nuScenes [45] according to the source of the annotated data. Both datasets inherit the data format of nuScenes, containing 700 training scenes and 150 validation scenes, with annotations for 17 categories. The evaluation range for OpenOccupancy is [-51.2 m, 51.2 m] in the X and Y directions, and [-3 m, 5 m] in the Z direction, using a voxel resolution of 0.2 meters. In comparison, the data coverage for Occ3D-nuScenes is [-40 m, 40 m] in the X and Y directions, and [-1 m, 5.4 m] in the Z direction, with

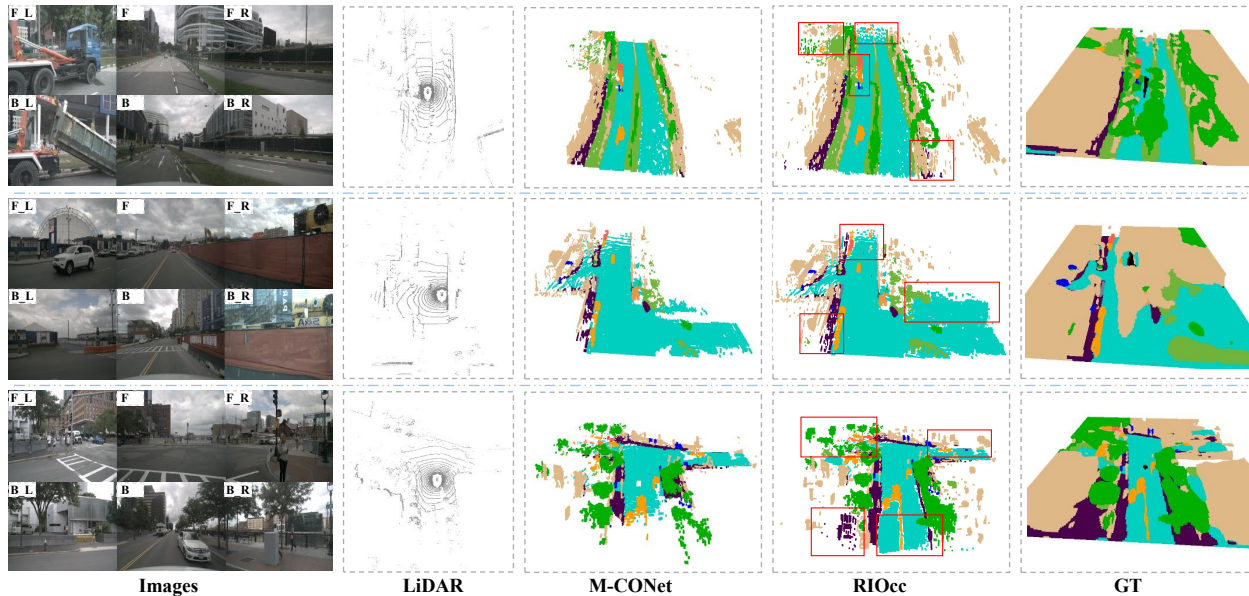


Figure 6. **The additional qualitative comparisons results between RIOcc and M-CONet.** The red box highlights the effectiveness in dealing with distant and occluded targets.

discretization using voxels of size [0.4 m, 0.4 m, 0.4 m].

**Metrics.** We adopt the official evaluation metrics, including IoU and mIoU.

## 4.2. Implementation Details

Our method is implemented based on MMDetection3D [8]. For the camera branch, we use ResNet50 pretrained on ImageNet as the image backbone, and the input image size is cropped to 256×704. For the LiDAR branch, we voxelize 10 LiDAR sweeps and employ a voxel encoder for the nuScenes dataset. During training, we use the AdamW optimizer, set the weight decay to 0.01, and an initial learning rate of  $1e^{-4}$ , with a multi-step learning rate scheduler for optimization. Training is conducted on four NVIDIA 3090 GPUs with a batch size of 4, for a total of 24 epochs.

## 4.3. Comparison with State-of-the-Art Methods

To ensure fair comparison, all results are either provided by the original authors or reproduced using the official code. We first compare RIOcc with some classical and advanced methods on Occ3D-nuScenes. The results are shown in Table 1, RIOcc achieves the latest state-of-the-art performance with an impressive 54.20 mIoU, outperforming RadOcc by 4.82 mIoU. To further evaluate the effectiveness of our proposed framework, we conduct additional performance tests on the nuScenes-Occupancy validation set. As shown in Table 2, RIOcc again achieves the best performance, with 35.4 IoU and 25.9 mIoU.

## 4.4. Ablation Studies

We conduct ablation experiments using the camera mask on the Occ3D-nuScenes dataset to evaluate the contribution of

each module to the overall framework.

**Effects of Downsampling Layers.** We conduct extensive experiments to evaluate the number of downsampling layers. Notably, all experiments are performed without the Wavelet Encoder and Semantic Encoder. As shown in the Table 3, when the number of downsampling layers is set to 0 (no downsampling), the model achieves an mIoU of 48.21, with a memory usage of 5.02 GB. As the number of downsampling layers increases, the mIoU continues to grow by 3.82 at 2 layers, with memory usage increasing slightly to 5.58 GB. However, when further downsampling to 3 layers, the performance improvement tends to plateau. Therefore, we choose to use 2 downsampling layers, which strikes a reasonable balance between computational load and significant performance improvement.

**Influences of Wavelet and Semantic Encoders.** To validate the effectiveness of the Wavelet Encoder and Semantic Encoder, we present an ablation study shown in Table 4. The results indicate that adding either the Wavelet Encoder or the Semantic Encoder alone can improve model performance. When both are incorporated, the mIoU increases from the baseline of 51.96 to 54.21. It demonstrates that introducing the Wavelet Encoder and Semantic Encoder effectively enhances the model’s ability to capture refined features and understand semantic information.

**Effectiveness of DBP.** As illustrated in the Table 5, we explore the contributions of the submodules within the DBP module. It obtains 0.63 mIOU with channel-wise alone and 0.71 mIOU with grid-wise alone. When both branches are combined to form the complete DBP module, the mIoU increases by 1.14. It indicates that the synergy between the

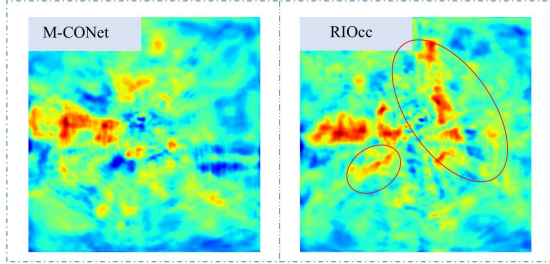


Figure 7. Features heatmaps after adaptive fusion.

Downsampling Layer	0	1	2	3
mIoU	48.21	51.12	52.03	52.15
Memory(GB)	5.02	5.32	5.58	5.82

Table 3. Ablation study of downsampling layer.

#	Semantic	mIoU
w/ Wavelet	✓	53.32 <b>54.21</b>
w/o Wavelet	✓	51.96 53.37

Table 4. Ablation Study of Wavelet and Semantic Encoders. It is worth noting that the default downsampling operation is used in the absence of wavelet or semantic encoding.

Channel-wise and Grid-wise branches has a greater advantage in capturing multi-level feature information.

**Impacts of Aggregation Region Size.** We illustrate the results to demonstrate that different region sizes in the DDA module help identify the optimal local aggregation size for enhancing feature representation and fusion. To ensure consistency in the number of final tokens, we uniformly sampling of the same number of anchors and test 4 different region sizes.  $1 \times 1$  means no feature aggregation, with each token corresponding to a single feature point. As indicated in Table 6, when the region size is set to  $5 \times 5$ , the model achieves the highest mIoU at 48.21. This setting proves an ability to preserve sufficient specificity while enhancing detection accuracy for large areas and distant objects. In contrast, when the region size is further increased to  $7 \times 7$ , detection accuracy slightly decreases. This result may be attributed to an overly large aggregation region, which smooths the feature maps and blurs local details.

**Comparison of Interaction Methods.** In Table 7, we compare the effectiveness of conventional BEV feature fusion methods with our proposed DDA module. The conventional addition and concatenation fusion strategies achieve mIoU of 46.58 and 46.69, respectively. In contrast, the DDA module results in a significant improvement, with the mIoU gaining about 1.57. The DDA module demonstrates a stronger impact on feature fusion, effectively enhancing

#	Channel Wised	Grid Wised	mIoU
1			53.07
2	✓		53.70
3		✓	53.78
4	✓	✓	<b>54.21</b>

Table 5. Ablation Study on Channel-Wised and Grid-Wised Operations. To reduce training time, this experiment uses only a single-scale feature for training.

region	$1 \times 1$	$3 \times 3$	$5 \times 5$	$7 \times 7$
mIoU	47.43	47.92	<b>48.21</b>	48.16

Table 6. Ablation Study of Aggregation Region Size.

#	Strategy	mIoU
1	Addition	46.58
2	Concatenation	46.69
3	DDA	<b>48.21</b>

Table 7. Ablation study of the Dual-BEV fusion strategy.

representation and improving scene understanding.

**Feature Alignment on Heatmaps.** To demonstrate that our model effectively enhances feature alignment in the LiDAR-camera fusion process, we present the feature heatmaps after adaptive fusion for both M-CONet and our method, as shown in the Figure 7. Note that we only select the first channel of the sample for visualization.

## 4.5. Visualization

As shown in Figure 6, we present the visualization of RIOcc on Occ3D-nuScenes without the camera mask. The results indicate that our RIOcc provides a more comprehensive prediction of the scene. It demonstrates improved detection performance for distant and occluded objects, along with a finer-grained occupancy prediction of the scene.

## 5. Conclusion

We propose RIOcc, a novel multi-modal 3D semantic occupancy prediction method, which is equipped with advanced BEV features refinement and interaction mechanisms. RIOcc integrates multi-modal data in a unified BEV space to achieve less computational burden. The LiDAR and camera branches extract refined BEV features independently, while the interaction branch effectively mitigates the disparity between multi-modal BEV features. Finally, the fused BEV features are transformed into voxel representation for occupancy prediction. Extensive experiments on the Occ3D-nuScenes and nuScenes-Occupancy datasets demonstrate the superiority of RIOcc over existing approaches. We believe our model will inspire new insights in the field.

## Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 62473205, Natural Science Foundation of Jiangsu Province No. BK20241893, and sponsored by Yong Academic Leader of Qing Lan Project in Jiangsu Province.

## References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 1
- [2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2
- [3] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Self-supervised open vocabulary occupancy estimation via volume rendering. *arXiv preprint arXiv:2407.17310*, 2024. 6
- [4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6
- [5] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2, 6
- [6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1
- [7] Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Zhaoqing Wang, Yandong Guo, and Shanghang Zhang. Bev-san: Accurate bev 3d object detection via slice attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17461–17470, 2023. 2
- [8] MMDetection3D Contributors. MMDetection3D: Open-MMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020. 7
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2
- [10] Michael Firman, Oisín Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5431–5440, 2016. 2
- [11] Yu Hong, Qian Liu, Huayuan Cheng, Danjiao Ma, Hang Dai, Yu Wang, Guangzhi Cao, and Yong Ding. Univision: A unified framework for vision-centric 3d perception. *arXiv preprint arXiv:2401.06994*, 2024. 6
- [12] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 fourth international conference on 3D vision (3DV)*, pages 92–101. Ieee, 2016. 2
- [13] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2
- [14] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 2, 6
- [15] Qi Jiang, Hao Sun, and Xi Zhang. Semanticbev-fusion: Rethink lidar-camera fusion in unified bird’s-eye view representation for 3d object detection. *arXiv preprint arXiv:2212.04675*, 2022. 1
- [16] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2561–2569, 2024. 2
- [17] Zheng Jiang, Jinqing Zhang, Yanan Zhang, Qingjie Liu, Zhenghui Hu, Baohui Wang, and Yunhong Wang. Fsd-bev: Foreground self-distillation for multi-view 3d object detection. *arXiv preprint arXiv:2407.10135*, 2024. 2
- [18] Yang Jiao, Zequn Jie, Shaoxiang Chen, Lechao Cheng, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2598–2606, 2024. 2
- [19] Alex H Lang, Sourabh Vora, Holger Caesar, Luming Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1
- [20] Heng Li, Yuenan Hou, Xiaohan Xing, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. *arXiv preprint arXiv:2408.09859*, 2024. 3, 6
- [21] Peidong Li, Wancheng Shen, Qihao Huang, and Dixiao Cui. Dualbev: Cnn is all you need in view transformation. *arXiv preprint arXiv:2403.05402*, 2024. 2
- [22] Yin hao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1, 2
- [23] Zhiqi Li, Wenhao Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: *Conference on Computer Vision and Pattern Recognition*, pages 5431–5440, 2016. 2

- Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1
- [24] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 2, 6
- [25] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 1, 3
- [26] Zhiwei Lin, Hongbo Jin, Yongtao Wang, Yufei Wei, and Nan Dong. Teocc: Radar-camera multi-modal occupancy prediction via temporal enhancement. In *ECAI 2024*, pages 129–136. IOS Press, 2024. 6
- [27] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2, 1
- [28] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. *arXiv preprint arXiv:2312.03774*, 2023. 6
- [29] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19936–19945, 2024. 6
- [30] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: Multi-sensor fusion framework for 3d semantic occupancy prediction. *IEEE Transactions on Intelligent Vehicles*, 2024. 3, 6
- [31] Jingyi Pan, Zipeng Wang, and Lin Wang. Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 2024. 3, 6
- [32] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024. 6
- [33] Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, and Deng Cai. Learning occupancy for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10281–10292, 2024. 2
- [34] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2
- [35] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 6
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4
- [37] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 1
- [38] Yining Shi, Kun Jiang, Ke Wang, Kangan Qian, Yunlong Wang, Jiushi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, and Diange Yang. Effocc: A minimal baseline for efficient fusion-based 3d occupancy network. *arXiv preprint arXiv:2406.07042*, 2024. 6
- [39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2
- [40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2
- [41] Yang Song and Lin Wang. Bico-fusion: Bidirectional complementary lidar-camera fusion for semantic and spatial-aware 3d object detection. *arXiv preprint arXiv:2406.19048*, 2024. 1
- [42] Ziyang Song, Guoxing Zhang, Lin Liu, Lei Yang, Shaoqing Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Robofusion: Towards robust multi-modal 3d object detection via sam. *arXiv preprint arXiv:2401.03907*, 2024. 1
- [43] Xin Tan, Wenbin Wu, Zhiwei Zhang, Chaojie Fan, Yong Peng, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision. *arXiv preprint arXiv:2405.10591*, 2024. 6
- [44] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xianguan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024. 6
- [45] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 6
- [46] Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xianguan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2404.15014*, 2024. 3, 6

- [47] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 3, 6, 2
- [48] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17158–17168, 2024. 2, 6
- [49] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 2
- [50] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 1, 6
- [51] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020. 1
- [52] Zhen Yang, Yanpeng Dong, and Heng Wang. Daocc: 3d object detection assisted multi-sensor fusion for 3d occupancy prediction. *arXiv preprint arXiv:2409.19972*, 2024. 6
- [53] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 5
- [54] Haiming Zhang, Xu Yan, Dongfeng Bai, Jiantao Gao, Pan Wang, Bingbing Liu, Shuguang Cui, and Zhen Li. Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7060–7068, 2024. 6
- [55] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird’s-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3348–3357, 2023. 2
- [56] Ji Zhang, Yiran Ding, and Zixin Liu. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction. *arXiv preprint arXiv:2403.05329*, 2024. 6
- [57] Jinqing Zhang, Yanan Zhang, Yunlong Qi, Zehua Fu, Qingjie Liu, and Yunhong Wang. Geobev: Learning geometric bev representation for multi-view 3d object detection. *arXiv preprint arXiv:2409.01816*, 2024. 2
- [58] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 6
- [59] Xiao Zhao, Xukun Zhang, Dingkan Yang, Mingyang Sun, Mingcheng Li, Shunli Wang, and Lihua Zhang. Maskbev: Towards a unified framework for bev detection and map segmentation. *arXiv preprint arXiv:2408.09122*, 2024. 1, 2
- [60] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023. 6