DGA-OCC: DISTANCE-GUIDED AND GATED-ADAPTIVE FUSION FOR MULTIMODAL 3D SEMANTIC OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing multimodal 3D semantic occupancy prediction methods typically perform feature fusion only in local voxel space or global BEV space, failing to account for cross-modal spatial biases and neglecting the complementary advantages of global and local collaborative fusion. To address this, this paper proposes a global-local collaborative multimodal fusion framework based on distance guidance and gating adaptation. Specifically, the gating-adaptive local fusion(GALF) achieves reliable local feature interaction in geometrically aligned voxel space through bidirectional cross-modal nearest neighbor association and learnable gating modulation; the dual-geometric interactive fusion(DGIF) combines cross-modal interaction alignment, camera-LiDAR dual-source geometric priors, and distance-aware attention to achieve spatially adaptive global fusion. Furthermore, the Mamba Split Block jointly models global structure and local details through frequency-aware decomposition, and the local-global fusion module(LGFM) further promotes bidirectional collaboration between camera BEV semantics and LiDAR 3D geometry. Experimental results show that the proposed method achieves mIoU of 54.46 and 25.9 on the Occ3D-nuScenes and nuScenes-Occupancy datasets, respectively, validating the effectiveness of the proposed framework and its components.

1 INTRODUCTION

3D semantic occupancy prediction is fundamental to autonomous driving perception, providing dense spatial and semantic understanding of the environment. By combining LiDAR's precise geometry with the rich semantics of cameras, multimodal fusion has become a dominant solution. However, current mainstream research on multimodal feature fusion still has significant limitations. Most methods simply emphasize feature interaction in global space (as shown in Figure 1(a)) or feature fusion at the local level (as shown in Figure 1(b)). Existing methods insufficiently exploit global–local complementarity: local fusion is vulnerable to calibration errors and empty voxels, while global fusion often relies on fixed weights that ignore spatially varying sensor reliability, resulting in coarse cross-modal interaction and limited scene adaptability.

To address the aforementioned issues, this paper proposes a global-local collaborative multimodal fusion framework that balances explicit local interaction with dynamic global aggregation, enabling hierarchical cross-modal collaboration in both local voxel space and global BEV space. Specifically, this paper designs a gated adaptive local fusion module (GALF) to perform bidirectional cross-modal K-nearest neighbor retrieval on camera and LiDAR non-empty voxels in a unified 3D voxel space, establishing neighborhood correspondences with local geometric consistency. Subsequently, a learnable gating network is used to adaptively adjust the cross-modal neighborhood response, thereby mitigating feature matching bias and insufficient semantic information caused by extrinsic parameter calibration errors and sparse empty voxels. To further model cross-modal complementary relationships globally, this paper proposes a dual-geometry interactive fusion module (DGIF). This module first maps the BEV features of the camera and LiDAR to a unified token space through the Cross-Modal Interaction Alignment (CMIA) module, and establishes bidirectional information interaction using multi-head self-attention. Then, it constructs a dual-source geometric prior by combining camera depth information and LiDAR explicit geometric statistics, and adaptively adjusts the cross-modal information aggregation method based on the geometric features and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

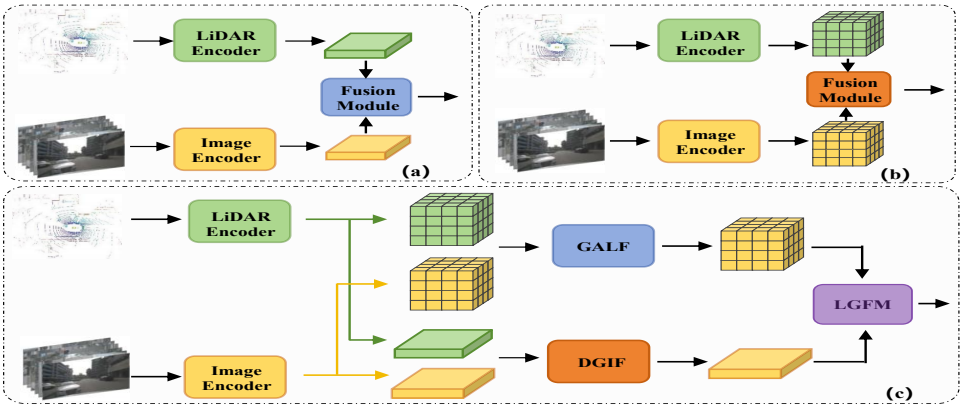


Figure 1: Schematic comparison of previous methods and the proposed model.

observation distance of different spatial regions by combining distance encoding and cross-attention. In the feature modeling stage, this paper introduces the Mamba Split Block, which divides the input features into low-frequency components emphasizing global structure and high-frequency components emphasizing local changes through frequency-aware decomposition. Mamba is used to model long-range spatial dependencies, convolution is used to extract edge and texture details, and residual connections are used to obtain a BEV representation that combines global structure and local details. Finally, this paper constructs a Local-Global Fusion Module (LGFM). First, it uses EfficientViT-Block and BottleNeckASPP to enhance the long-range dependence and multi-scale context of camera BEV features, while compressing LiDAR 3D voxel features along the height dimension into the BEV space. On this basis, through adaptive modal weighting and bidirectional cross-attention, the camera semantic features are made aware of the LiDAR geometry, and camera semantic information is added to the LiDAR geometric features, further realizing the collaborative fusion of local geometric representation and global semantic context.

Overall, the contributions of this paper are listed in three aspects as follows:

- We propose a global-local collaborative multimodal fusion framework for 3D semantic occupancy prediction. This framework jointly utilizes fine-grained geometric information provided by LiDAR and rich semantic information captured by multi-view images, thereby achieving complementary integration of local geometric fusion and global semantic fusion.
- We design two complementary fusion modules, GALF and DGIF. GALF achieves adaptive local fusion through bidirectional neighborhood association and gated modulation, while DGIF realizes dynamic global fusion through cross-modal interaction alignment, dual-source geometric priors, and distance-guided attention.
- We introduce the Mamba Split Block and LGFM to enhance the modeling of global structures and local details, and to promote collaborative interaction between camera BEV semantics and LiDAR 3D geometry, respectively. Experimental results show that the proposed method achieves 54.46 and 25.9 mIoU on Occ3D-nuScenes and nuScenes-Occupancy, respectively.

2 Related Work

BEV-based 3D Perception. In recent years, bird's-eye view (BEV) perception has emerged as an indispensable technology in autonomous driving by virtue of its global field-of-view and high computational efficiency. Driven by the decreasing cost of cameras, vision-only 3D object detection has achieved rapid progress. LSS Philion & Fidler (2020) stands as a representative work in this domain, which leverages geometric modeling to construct depth maps. BEVDet Huang et al. (2021) first incorporated this paradigm into a complete 3D detection framework. Subsequent studies further advanced this line of research: BEVDepth Li et al. (2023a) introduced explicit depth supervision optimization.

BEVFusion Liu et al. (2023) projects features from different sensors into the BEV plane to implement feature fusion in a unified manner. On this basis, MaskBEV Zhao et al. (2024) additionally

introduces task-agnostic masks into the BEV feature space. Despite the promising performance of existing approaches, they still encounter difficulties in detecting out-of-distribution or non-target objects.

Camera-based 3D Semantic Occupancy Prediction. To address the inherent drawbacks of conventional object detection methods, 3D semantic occupancy prediction has attracted widespread research attention. MonoScene Cao & De Charette (2022) adopts optics-inspired 2D-to-3D feature projection to infer dense geometric and semantic scene information. TPVFormer Huang et al. (2023) integrates a three-view representation with Transformer architectures to map 2D image features into 3D space for occupancy prediction. More recently, OccupancyM3D Peng et al. (2024) directly learns 3D occupancy states and embeds occupancy cues into the feature representation pipeline of monocular 3D detection. PanoOcc Wang et al. (2024b) utilizes voxel queries alongside a coarse-to-fine optimization strategy, enabling effective semantic modeling of all objects within complex driving scenarios. Even so, purely vision-based occupancy prediction suffers from inaccurate depth estimation and inadequate geometric constraints. Against this backdrop, multimodal perception has become a more robust and dependable alternative.

multimodal 3D Semantic Occupancy Prediction. To further boost the prediction accuracy of 3D semantic occupancy forecasting, research efforts have increasingly shifted toward the exploration of multimodal perception solutions. OpenOccupancy Wang et al. (2023) puts forward the pioneering benchmark framework for LiDAR-camera based semantic occupancy prediction. On the basis of this work, OccGen Wang et al. (2024a) develops a novel noise-to-occupancy generative paradigm, which eliminates random Gaussian noise in an iterative manner to generate and refine high-quality 3D semantic occupancy maps. Co-Occ Pan et al. (2024a) establishes a multimodal 3D semantic occupancy prediction architecture that integrates explicit feature fusion with volume rendering regularization.

OccMamba Li et al. (2025) incorporates the Mamba structure together with a 3D-to-1D reordering strategy to improve the computational efficiency of large-scale 3D voxel processing. Although multimodal perception techniques have achieved remarkable progress, they still remain plagued by unresolved issues including insufficient feature interaction, heavy computational overhead, and rough feature representation, leaving ample room for further optimization and improvement.

3 METHODS

The overall framework is illustrated in Figure 2. Multi-view images and LiDAR point clouds are first encoded into camera BEV features and LiDAR voxel features. The camera branch is enhanced by LSS and the Mamba Split Block, then reconstructed into 3D features. GALF performs local cross-modal fusion in voxel space, while DGIF conducts distance-guided global fusion in BEV space. Finally, LGFM integrates the fused representations for semantic occupancy prediction.

3.1 MAMBA SPLIT BLOCK

We introduce a lightweight Mamba Split Block to preserve fine-grained geometric details while modeling global context in BEV space. Given an input feature F, a 1×1 convolution and a learnable partition coefficient decompose it into low- and high-frequency components:

$$F_{low} = \alpha F, F_{high} = (1 - \alpha)F \tag{1}$$

The low-frequency branch first applies a 7×7 convolution and average pooling to aggregate contextual information and reduce token redundancy. A two-dimensional selective scan then models long-range spatial dependencies with linear complexity:

$$G = \text{Linear}\left(SS2D\left(AP\left(\text{Conv}_{7\times7}\left(F_{low}\right)\right)\right)\right) \tag{2}$$

To recover local variations suppressed during global modeling, G is upsampled and subtracted from the low-frequency input:

$$R_{\text{high}} = F_{\text{low}} - UP(G) \tag{3}$$

The resulting residual is combined with the high-frequency branch and refined by a 3×3 convolution to enhance object boundaries and texture details:

$$H = \text{Conv}_{3\times3}\left(R_{high} + F_{high}\right) \tag{4}$$

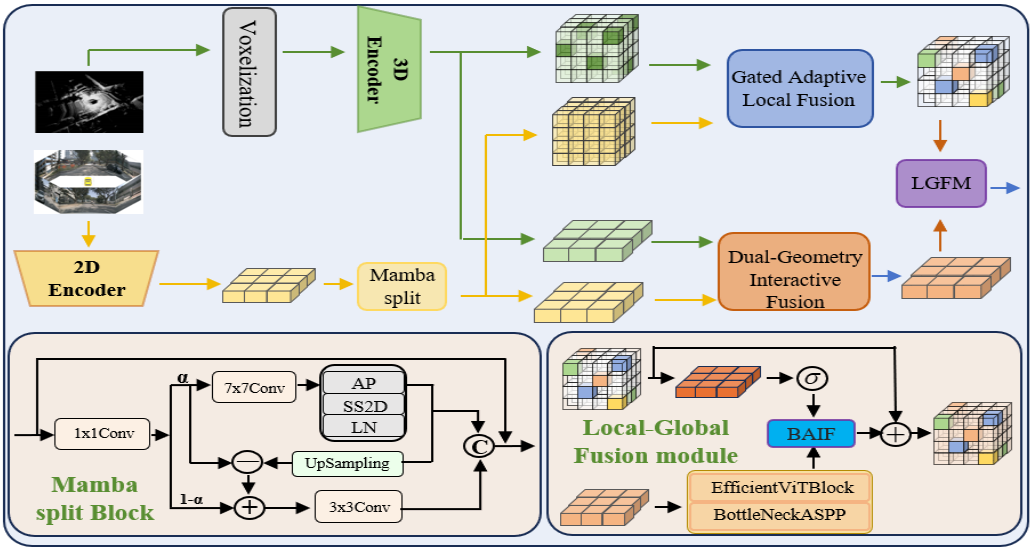


Figure 2: Overall architecture of the proposed method. Multi-view images and LiDAR point clouds are encoded into semantic and geometric features, with camera features enhanced by the Mamba Split Block. GALF performs gated local fusion in 3D voxel space, while DGIF conducts distance-aware global fusion in BEV space using cross-modal alignment and dual-source geometric priors. LGFM then integrates camera BEV semantics with LiDAR geometry for final 3D semantic occupancy prediction. The internal designs of the Mamba Split Block and LGFM are shown at the bottom.

Finally, the low-frequency global features modeled by Mamba, the high-frequency detail features refined by convolution, and the shortcut residual of the original input are fused together to obtain the output feature F_{img}^{BEV} with clear local details.

3.2 Gated Adaptive Local Fusion

Imprecise external parameter calibration between LiDAR and camera makes the 2D-to-3D geometric view transformation fail to accurately map image features to corresponding positions in 3D space. A large number of empty voxels also exist in LiDAR data. These problems hinder effective interaction between LiDAR and camera features, and further lead to inaccurate feature fusion and semantic information loss in multimodal fusion. To solve the above problems, this paper proposes the GALF module as shown in Figure 3. On the basis of geometric alignment, this module enables LiDAR features to fully absorb rich semantic information from camera data. It also alleviates fusion errors caused by external parameter calibration, and improves the semantic representation capability of sparse LiDAR features.

Specifically, multi-view image features first perform the Lift-Splat operation via view transformation and are mapped to a unified dense 3D voxel grid space to generate camera voxel feature $F_{img} \in R^{B\times C\times H\times W\times L}$. Meanwhile, LiDAR point clouds are encoded by a sparse voxel encoder and processed into the same 3D voxel grid space to obtain $F_{lidar} \in R^{B\times C\times H\times W\times L}$. Among them, B represents the batch size, C denotes the number of feature channels, and H, W, L correspond to the height, width and depth dimensions of the voxel grid respectively. To avoid redundant calculation of the complete 3D voxel grid, subsequent operations are only performed on non-empty voxels with valid features. The intensity map in the shape of $R^{H\times W\times L}$ is obtained by summing F_{img} and F_{lidar} along the channel dimension. The sum value at each position in the intensity map directly indicates whether the corresponding voxel has valid activated features. A non-zero sum means the voxel contains valid features; otherwise, it is an empty voxel. By traversing the intensity map, all discrete index coordinates of voxels with non-zero sum values are accurately screened out. Finally, the camera non-empty voxel coordinate set $P_{img} \in R^{N_{img}\times 3}$ and the LiDAR non-empty voxel coordinate set $P_{lidar} \in R^{ptr\times 3}$ are obtained, and each row of coordinates is expressed in the form of (h,w,l). Then the corresponding non-empty valid features are accurately extracted from the original F_{img} and F_{lidar} according to these coordinates, acquiring $X_{img} \in R^{N_{img}\times 3}$ and $X_{lidar} \in R^{N_{ptr}\times 3}$.

Next, bidirectional cross-modal K-nearest neighbor neighborhood retrieval and feature matching are implemented. On the one hand, taking LiDAR non-empty voxel coordinates P_{lidar} as queries and camera non-empty voxel coordinates P_{img} as keys, farthest point sampling is first performed on the query points to obtain representative feature points. The Euclidean distances between representative points and all keys are calculated. Ball query is used to screen neighboring key points within a preset geometric radius. The neighboring key point features of representative points are back-propagated to all covered query points, and each LiDAR non-empty voxel is matched with K geometrically adjacent camera neighborhood features. On the other hand, taking camera non-empty voxel coordinates P_{img} as query points and LiDAR non-empty voxel coordinates P_{lidar} as key points, the same K-nearest neighbor retrieval process is executed to establish bidirectional geometric neighborhood association between camera and LiDAR. Afterwards, it enters the learnable gated weighted feature enhancement stage. The retrieved K cross-modal neighborhood features are concatenated along the channel dimension and input into the learnable gated network.

The concatenated features are compressed to the same channel dimension as the original modality. Element-wise multiplication is then adopted to realize adaptive modulation of semantic weights. Specifically, the LiDAR side generates adaptive semantic weights ω_{lidar} through the gated network. The camera neighborhood features are weighted by these weights, and then fused with the LiDAR non-empty features X_{lidar} via element-wise multiplication to obtain the semantically enhanced LiDAR non-empty features X'_{lidar}. On the camera side, the same gated network generates semantic weights ω_{img}. The LiDAR neighborhood features are weighted and fused with the camera non-empty features X_{img} by element-wise multiplication, yielding the semantically enhanced camera non-empty features X'_{img}. The mathematical formulation is defined as follows:

$$\begin{aligned} X'_{\text{lidar}} &= X_{\text{lidar}} \odot \left(\omega_{\text{lidar}} \cdot X_{\text{img}}^{\text{neighbor}}\right) \\ X'_{\text{img}} &= X_{\text{img}} \odot \left(\omega_{\text{img}} \cdot X_{\text{lidar}}^{\text{neighbor}}\right) \end{aligned} \tag{5}$$

Among them, $X_{img}^{neighbor}$ denotes the set of camera neighborhood features, $X_{lidar}^{neighbor}$ represents the set of LiDAR neighborhood features, and $\odot$ stands for the element-wise multiplication operation.

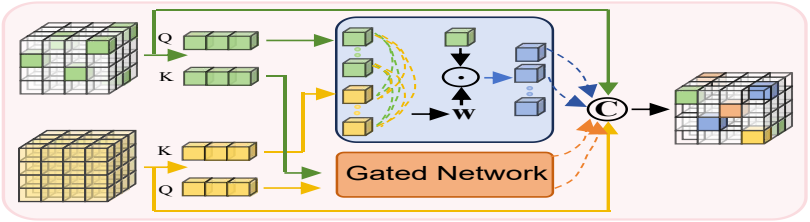


Figure 3: Framework of the Gated Adaptive Local Fusion (GALF) module.

After feature enhancement, X'_{img} and X'_{lidar} are back-projected to the corresponding non-empty coordinate positions of the original dense 3D voxel grid respectively. The LiDAR fused voxel feature F_{lidar_fuse} and camera fused voxel feature F_{img_fuse} are generated accordingly, while the remaining empty voxel positions remain zero. This ensures that the spatial dimension of the fused features is completely consistent with that of the original voxel grid. Finally, multimodal feature fusion and 3D context modeling are performed. F_{img}, F_{lidar}, F_{img_fuse}, F_{lidar_fuse} are concatenated along the channel dimension into a feature tensor with 4C channels. Two stacked 3D convolution modules together with BN layers are adopted in sequence to complete feature normalization, channel compression and local 3D context modeling. The unified 3D voxel feature F_{fuse}^{3D} after multimodal fusion is finally obtained, which is formulated as follows:

$$F_{\text{concat}} = \text{Concat}\left(F_{\text{img}}, F_{\text{lidar}}, F_{\text{lidar_fuse}}, F_{\text{img_fuse}}\right) \tag{6}$$

$$F_{\text{fuse}}^{\text{3D}} = \text{BN}\left(\text{Conv3D}_2\left(\text{ReLU}\left(\text{BN}\left(\text{Conv3D}_1\left(F_{\text{concat}}\right)\right)\right)\right)\right) \tag{7}$$

3.3 Dual-Geometry Interactive Fusion

To fully leverage the complementary advantages of Camera and LiDAR sensors in terms of semantic and geometric information, this paper proposes a DGIF. This module achieves effective fusion of Camera-LiDAR BEV features through cross-modal interactive alignment, dual-source geometric prior modeling, and distance-aware attention fusion. Since Camera BEV and LiDAR BEV originate from different sensors, their feature distributions and information focuses differ. Direct feature fusion can easily lead to inconsistent cross-modal responses. Therefore, this paper designs a Cross-Modal Interaction Alignment (CMIA) module to establish explicit information interaction between the two modalities before fusion. Specifically, the Camera BEV features and LiDAR BEV features are first mapped to a unified feature space using lightweight convolutions and then flattened into BEV tokens. Subsequently, the two types of tokens are concatenated, and spatial location encoding and modality encoding are added to preserve the spatial structure of the BEV and sensor origin information. The fused token is input into a multi-head self-attention module (CMIA), which globally models the relationships between different modalities to achieve bidirectional information exchange between the camera and LiDAR. After CMIA, the enhanced BEV features are obtained.To enhance the spatial constraints during the fusion process, a dual-source geometric prior is constructed, utilizing both camera depth and LiDAR geometry information. For the Camera branch, a Camera Geometry Prior is generated using multi-view depth estimation results. First, the depth information is projected into the BEV space, and then a dense spatial depth representation is obtained through convolutional encoding. For the LiDAR branch, this paper extracts explicit geometric information from the original point cloud to construct a LiDAR Geometry Prior. This includes statistical features such as distance distribution, height information, point cloud density, and occupancy status within the BEV mesh. Compared to directly using LiDAR BEV features, this design avoids functional overlap between semantic features and geometric priors, while providing more explicit spatial constraints. Subsequently, the Camera Geometry Prior and the LiDAR Geometry Prior are fused to obtain the dual-source geometric prior. After obtaining the interactively enhanced BEV features and dual-source geometric priors, DGIF further employs a distance-guided attention mechanism to complete global fusion. First, a distance encoding is constructed based on the BEV's spatial location, and distance-aware geometric features are generated by combining dual-source geometric priors. This process enables the model to adaptively adjust the fusion strategy for different regions based on the spatial distance between the target and the vehicle. Subsequently, the Camera BEV features are used as the query branch. Simultaneously, the LiDAR BEV features and the distance-enhanced geometric prior are jointly mapped as keys and values. Cross-Attention is then used to aggregate the two modalities to obtain the fused BEV representation.

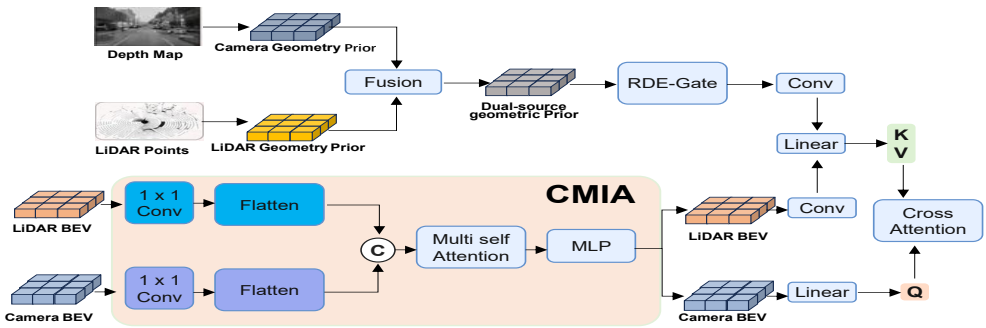


Figure 4: Framework of the Dual-Geometry Interactive Fusion (DGIF) module.

3.4 Local-Global Fusion Module

To enhance the collaborative representation of LiDAR geometric features and camera semantic features in the local-global fusion stage, this paper designs a LGFM. This module takes camera BEV features and LiDAR 3D voxel features as input, performs cross-modal bidirectional interaction in a unified BEV space, and remaps the fused BEV representation back to the 3D voxel space for subsequent semantic occupancy prediction. In the overall framework, image features are processed by LSS and Mamba Split Block to obtain BEV representations, and LiDAR features are processed by voxel encoding to obtain 3D geometric representations, which then proceed to the local-global fusion stage.

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	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
FB-OccLi et al. (2023b)	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*Hong et al. (2024)	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*Tan et al. (2025)	C	256×704	R50	43.64	14.29	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
EFFOcc*Shi et al. (2025)	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
RenderOccPan et al. (2024b)	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
CTF-OccTian et al. (2023)	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*Wang et al. (2024b)	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*Lu et al. (2024)	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
SA-Occ*Chen et al. (2025)	C	512×1408	Swin-B	44.64	13.70	51.10	30.40	52.40	56.90	27.00	31.60	30.20	30.30	37.60	44.00	84.80	46.10	58.60	**62.10**	53.10	49.10
OccFusion*Zhang et al. (2024b)	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	44.65	57.13	60.01	62.46	61.25
RadOcc*Zhang et al. (2024a)	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
SDGOcc*Duan et al. (2025)	C&L	256×704	R50	51.66	13.21	57.77	24.30	60.33	64.28	36.21	39.44	52.36	35.80	**50.91**	53.65	84.56	47.45	58.00	61.61	**70.67**	67.65
Ours	C&L	256×704	R50	**54.46**	12.04	**59.96**	**36.84**	**62.49**	**68.98**	**36.74**	**47.67**	**58.57**	**44.52**	49.76	**54.69**	**85.47**	46.39	**58.61**	62.17	70.87	**69.97**

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

Method	Modality	Resolution	Image Backbone	LiDAR Backbone	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoSceneCao & De Charette (2022)	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-CONetWang et al. (2023)	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
SparseOccTang et al. (2024)	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
L-CONetWang et al. (2023)	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
M-CONetWang et al. (2023)	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-OccPan et al. (2024a)	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
OccGenWang et al. (2024a)	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
OccFusionMing et al. (2024)	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
EFFOccShi et al. (2025)	C&L	256×704	R18	VoxelNet	30.8	22.9	28.1	16.7	22.1	27.3	13.0	24.8	**36.2**	**22.6**	16.8	21.6	29.4	13.9	18.2	20.6	26.5	28.8
MS-OccWei et al. (2025)	C&L	900×1600	R50	VoxelNet	32.1	25.3	29.4	**23.8**	24.2	28.7	17.5	27.5	29.0	21.1	16.9	25.3	33.4	23.5	24.4	24.6	27.6	**29.4**
GaussianWorldZuo et al. (2025)	C&L	900×1600	R101	VoxelNet	33.4	22.1	21.3	14.1	**27.7**	**31.8**	13.7	17.4	13.7	11.5	15.1	24.0	**43.0**	24.9	**28.9**	**26.7**	15.7	24.7
OccMambaLi et al. (2025)	C&L	900×1600	R50	VoxelNet	33.7	25.1	29.6	20.2	25.7	28.5	16.7	25.0	23.2	19.9	**20.3**	24.5	36.1	25.3	25.1	24.8	**27.7**	28.9
Ours	C&L	256×704	R50	VoxelNet	**35.1**	**25.9**	**30.0**	19.6	25.6	28.5	**18.5**	24.7	31.7	21.7	19.6	**25.0**	37.4	24.5	25.8	25.1	27.1	28.9

Specifically, let the input camera BEV feature be F_C^{BEV}, and the LiDAR 3D voxel feature be F_L^{3D}. First, LGFM enhances the semantic context of the camera BEV branch. This paper uses EfficientViTBlock to model long-range spatial dependencies in the BEV plane and utilizes BottleNeckASPP to aggregate multi-scale contextual information, thereby obtaining the enhanced camera BEV feature $\tilde{F}_{\mathrm{C}}^{BEV}$.Meanwhile, to enable the 3D geometric information of LiDAR to interact with camera BEV features in the same space, compresses the LiDAR 3D voxel features along the height dimension into a LiDAR BEV representation $\tilde{F}_{\mathrm{L}}^{BEV}$. This operation transforms the spatial geometry of LiDAR to a BEV perspective, enabling it to be aligned and fused with camera BEV features. Subsequently, LGFM inputs $\tilde{F}_{\mathrm{C}}^{BEV}$ and $\tilde{F}_{\mathrm{L}}^{BEV}$ into Bidirectional Adaptive Interaction Fusion Module(BAIF). This unit first performs channel alignment through 1×1 convolution and generates adaptive fusion weights based on the concatenated features of the two modalities. These weights are used to dynamically adjust the importance of camera semantic information and LiDAR geometric information at different spatial locations. Then, the module constructs cross-modal interactions in two directions: on one hand, using camera BEV features as queries and LiDAR BEV features as keys and values, enabling camera semantic features to perceive LiDAR geometric structure; on the other hand, using LiDAR BEV features as queries and camera BEV features as keys and values, enabling LiDAR geometric features to obtain semantic supplementation. This bidirectional interaction method avoids the information bias caused by unidirectional modulation, allowing the two modalities to establish a more complete correspondence in the global BEV space.

4 Experiment

4.1 Dataset and Metrics

We conduct extensive experiments on benchmarks derived from the nuScenes dataset. According to the annotation sources, the evaluation is performed on two commonly used occupancy prediction benchmarks, namely nuScenes-Occupancy Caesar et al. (2020) and Occ3D-nuScenes Tian

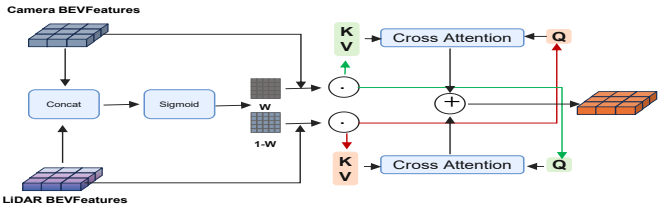


Figure 5: Framework of the Bidirectional Adaptive Interaction Fusion (BAIF) module.

et al. (2024). Both benchmarks follow the data organization of nuScenes and contain 700 training scenes and 150 validation scenes, with semantic annotations covering 17 categories. For nuScenes-Occupancy, the evaluation range is set to $[-51.2\,\mathrm{m}, 51.2\,\mathrm{m}]$ along the X and Y axes and $[-3\,\mathrm{m}, 5\,\mathrm{m}]$ along the Z axis, with a voxel resolution of $0.2\,\mathrm{m}$. In contrast, Occ3D-nuScenes covers the spatial range of $[-40\,\mathrm{m}, 40\,\mathrm{m}]$ along the X and Y axes and $[-1\,\mathrm{m}, 5.4\,\mathrm{m}]$ along the Z axis, and discretizes the 3D space into voxels with a size of $[0.4\,\mathrm{m}, 0.4\,\mathrm{m}, 0.4\,\mathrm{m}]$. We adopt the official evaluation metrics, including intersection over union (IoU) and mean intersection over union (mIoU), for quantitative evaluation.

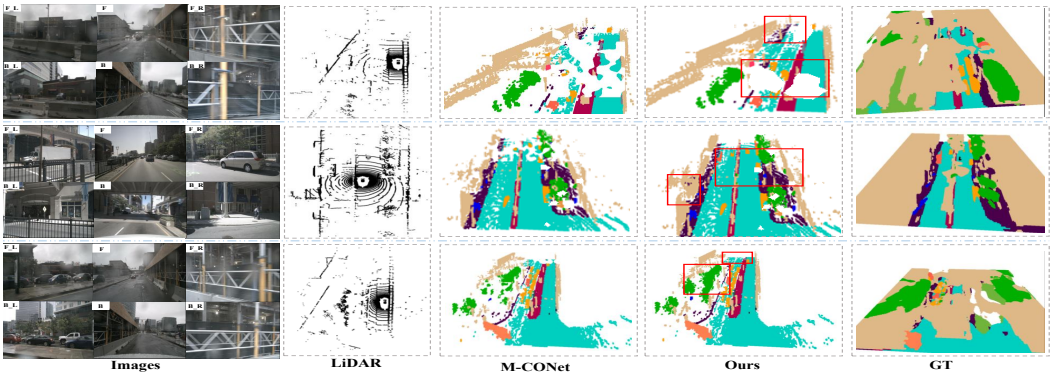


Figure 6: Comparative results of the proposed model on the nuScenes-Occupancy validation set, where red boxes highlight the effectiveness in handling distant and occluded objects.

4.2 Implementation Details

The method in this paper is implemented based on the MMDetection3D framework. For the camera branch, ImageNet pretrained ResNet50 is adopted as the image backbone network, and the input images are cropped to a resolution of 256×704. For the LiDAR branch, ten consecutive LiDAR sweeps are voxelized, and a voxel encoder adapted to the nuScenes dataset is utilized. In the training phase, the AdamW optimizer is employed with a weight decay of 0.01 and an initial learning rate of 1e-4, combined with a multistep learning rate scheduler for parameter optimization. The model is finally trained on four NVIDIA 3090 GPUs with a batch size of 4 and a total training epoch of 24.

4.3 Comparison with State-of-the-Art Methods

We conduct comparative experiments of the proposed model with various classic algorithms and advanced methods on the Occ3D-nuScenes dataset. As shown in Table 1, the model in this paper achieves state-of-the-art performance with an mIoU of 54.46, outperforming the SDGOcc method Duan et al. (2025) by 2.8 percentage points. We further evaluate the effectiveness of the proposed framework on the nuScenes-Occupancy validation set. The results in Table 2 demonstrate that our model still delivers excellent performance, reaching an IoU of 35.1 and an mIoU of 25.9 respectively.

4.4 Ablation Studies

Effectiveness of Mamba Split: To verify the contribution of the Mamba Split module to the model, corresponding ablation experiments are presented in Table 3. The experimental results show that simply introducing Mamba Split ($\alpha = 0.5$) improves the mIoU by 1.26. Compared with the wavelet

Table 3: Ablation study of Mamba Split

#	Wavelet	Mamba Split	mIoU
1			53.2
2	✓		54.13
3		✓	**54.46**

Table 4: Ablation study of GALF

#	L2C	C2L	mIoU
1			53.05
2	✓		53.76
3		✓	53.8
4	✓	✓	**54.46**

encoder, the mIoU is further increased by 0.33. It demonstrates that the introduction of Mamba Split effectively enhances the model's capability to capture local detail information.

Impact of GALF and Its Internal Components: To verify the effect of the GALF module and each of its internal components on the model, detailed ablation experiments on GALF are conducted respectively, with the results presented in Table 4. Compared with simple 3D feature concatenation, the introduction of GALF brings a 1.41 improvement in mIoU. Notably, when only using LiDAR non-empty voxel coordinates P_{lidar} as queries and camera non-empty voxel coordinates P_{img} as keys (C2L), the mIoU increases by 0.75; when taking camera non-empty voxel coordinates P_{img} as queries and LiDAR non-empty voxel coordinates P_{lidar} as keys (L2C), the mIoU rises by 0.71.

Table 5: Ablation study of DGIF versus traditional interaction methods

#	strategy	mIoU
1	Add	53.14
2	Cat.	53.21
3	DGIF	53.76
4	DGIF(CMIA)	**54.46**

Comparison between DGIF and Traditional Interaction Methods:To verify the advantages of the DGIF module over traditional feature interaction methods, we conduct corresponding ablation experiments in Table 5. The experimental results show that traditional feature addition (Add) and feature concatenation (Cat.) achieve mIoU of 53.14 and 53.21, respectively; when replacing them with the DGIF proposed in this paper, the mIoU increases to 53.76, and the DGIF (CMIA) module combined with CMIA further improves the mIoU to 54.46, demonstrating the significant gain of this improvement strategy on the model's feature interaction performance.

4.5 Visualization

As shown in Figure 6, the visualization results of the proposed model on nuScenes-Occupancy are presented without camera masks. The results demonstrate that our model yields more comprehensive scene prediction. It achieves improved detection performance for distant and occluded objects, as well as finer-grained occupancy prediction for the overall scene.

5 Conclusion

This paper addresses the problems of unstable local geometric alignment, lack of spatial adaptability in global fusion, and easy loss of local details in multimodal 3D semantic occupancy prediction, and proposes a global-local collaborative multimodal fusion framework. This framework establishes bidirectional cross-modal neighborhood associations in 3D voxel space through GALF and achieves reliable local feature fusion using a gating mechanism. It achieves spatially adaptive global information aggregation through DGIF combined with cross-modal interactive alignment, camera-LiDAR dual-source geometric priors, and distance-aware attention. Simultaneously, Mamba Split Block models the global structure and local details respectively, and LGFM further promotes the bidirectional collaboration between camera BEV semantics and LiDAR 3D geometric information.

AI USE STATEMENT

The authors used Generative AI (GenAI) tools to assist with manuscript preparation and to correct minor grammatical errors in the manuscript. No experimental data or reported results were generated using GenAI tools. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Detailed implementation, training, evaluation, and rollout protocols are provided in the appendix to facilitate reproduction of the reported results.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

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*, pp. 3991–4001, 2022.

Chen Chen, Zhirui Wang, Taowei Sheng, Yi Jiang, Yundu Li, Peirui Cheng, Luning Zhang, Kaiqiang Chen, Yanfeng Hu, Xue Yang, and Xian Sun. SA-Occ: Satellite-assisted 3D occupancy prediction in real world. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27021–27030, 2025. doi: 10.1109/ICCV51701.2025.02508.

ZaiPeng Duan, ChenXu Dang, Xuzhong Hu, Pei An, Junfeng Ding, Jie Zhan, YunBiao Xu, and Jie Ma. SDGOCC: Semantic and depth-guided bird's-eye view transformation for 3D multimodal occupancy prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6751–6760, 2025. doi: 10.1109/CVPR52734.2025.00633.

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.

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.

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*, pp. 9223–9232, 2023.

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. OccMamba: Semantic occupancy prediction with state space models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11949–11959, 2025. doi: 10.1109/CVPR52734.2025.01116.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023a.

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*, pp. 6919–6928, 2023b.

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)*, pp. 2774–2781. IEEE, 2023.

Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. *Advances in Neural Information Processing Systems*, 37:79618–79641, 2024.

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.

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*, 2024a.

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)*, pp. 12404–12411. IEEE, 2024b.

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*, pp. 10281–10292, 2024.

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*, pp. 194–210. Springer, 2020.

Yining Shi, Kun Jiang, Jinyu Miao, Ke Wang, Kangan Qian, Yunlong Wang, Jiusi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, and Diange Yang. EFFOcc: Learning efficient occupancy networks from minimal labels for autonomous driving. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17008–17015, 2025. doi: 10.1109/IROS60139.2025.11247287.

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. *IEEE Transactions on Intelligent Transportation Systems*, 26(4):5613–5623, 2025. doi: 10.1109/TITS.2025.3539627.

Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan 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*, pp. 15035–15044, 2024.

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*, 36:64318–64330, 2023.

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.

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2404.15014*, 2024a.

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*, pp. 17850–17859, 2023.

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*, pp. 17158–17168, 2024b.

Zhiqiang Wei, Lianqing Zheng, Jianan Liu, Tao Huang, Qing-Long Han, Wenwen Zhang, and Fengdeng Zhang. Ms-Occ: Multi-stage lidar-camera fusion for 3D semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 11(1):370–377, 2025.

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*, volume 38, pp. 7060–7068, 2024a.

Ji Zhang, Yiran Ding, and Zixin Liu. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction. *arXiv preprint arXiv:2403.05329*, 2024b.

Xiao Zhao, Xukun Zhang, Dingkang 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.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3D occupancy prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6772–6781, 2025. doi: 10.1109/CVPR52734.2025.00635.

A Appendix

Table 6: Ablation study of LGFM.

#	ViTBlock	BASPP	BAIF	mIoU
1				53.4
2	✓			53.79
3		✓		53.67
4			✓	53.81
5	✓	✓	✓	**54.46**

Table 7: Ablation analysis of each module's contribution to the model

#	GALF	DGIF	LGFM	mIoU
1				50.79
2	✓			52.22
3	✓	✓		53.59
4	✓	✓	✓	**54.46**

Effectiveness of LGFM and Its Internal Components: As shown in the ablation experiments in Table 6, we take the baseline model without extra components for comparison. Introducing ViTBlock alone boosts the mIoU to 53.79; BASPP alone yields an mIoU of 53.67; BAIF alone achieves 53.81. When all three components are integrated into the complete LGFM module, the mIoU reaches 54.46, with a 1.06 improvement over the baseline. The results prove each sub-component of LGFM brings independent performance gains, and their combination produces obvious synergy to strengthen the model's feature representation capability.

Comprehensive Analysis of Experimental Results for Each Module: To improve the overall logicality and readability, the experimental results of GALF, DGIF, and LGFM are systematically organized and comprehensively compared in Table 7. By progressively adding GALF, DGIF, and LGFM one by one, the table quantitatively presents the individual performance gain of each module as well as the synergistic effect after their combination, providing valuable experimental references for future research.