SDGOCC: Semantic and Depth-Guided Bird's-Eye View Transformation for 3D Multimodal Occupancy Prediction

Anonymous CVPR submission

Paper ID 9848

Abstract

Multimodal 3D occupancy prediction has garnered significant attention for its potential in autonomous driving. However, most existing approaches are single-modality: camera-based methods lack depth information, while LiDAR-based methods struggle with occlusions. Current lightweight methods primarily rely on the Lift-Splat-Shoot (LSS) pipeline, which suffers from inaccurate depth estimation and fails to fully exploit the geometric and semantic information of 3D LiDAR points. Therefore, we propose a novel multimodal occupancy prediction network called SDG-OCC, which incorporates a joint semantic and depth-guided view transformation coupled with a fusion-to-occupancy-driven active distillation. The enhanced view transformation constructs accurate depth distributions by integrating pixel semantics and co-point depth through diffusion and bilinear discretization. The fusion-to-occupancy-driven active distillation extracts rich semantic information from multimodal data and selectively transfers knowledge to image features based on LiDAR-identified regions. Finally, for optimal performance, we introduce SDG-Fusion, which uses fusion alone, and SDG-KL, which integrates both fusion and distillation for faster inference. Our method achieves state-of-the-art (SOTA) performance with real-time processing on the occ3d-nuscenes dataset and shows comparable performance on the more challenging surround-nuscenes dataset, demonstrating its effectiveness and robustness. The code will be released at https://github.com/

1. Introduction

Accurate 3D perception of the surrounding environment forms the cornerstone of modern autonomous driving systems and robotics, ensuring efficient planning and safe control [6, 9]. In recent years, advancements in 3D object detection [13, 16, 19, 36, 39] and semantic segmentation [10, 11, 35, 37, 43] have significantly propelled the field of 3D perception. However, object detection relies on strict bounding boxes, making it difficult to recognize arbitrary shapes or unknown objects, while semantic segmentation struggles with fine-grained classification in complex scenes, especially under occlusion and overlap. In this context, 3D semantic occupancy prediction [28, 30] offers a more comprehensive approach to environment modeling. It simultaneously estimates the geometric structure and semantic categories of scene voxels, assigns labels to each 3D voxel, and provides a more complete perception, showing stronger robustness to arbitrary shapes and dynamic occlusions.

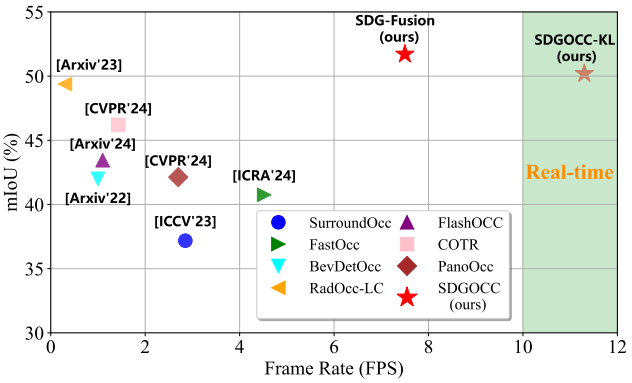
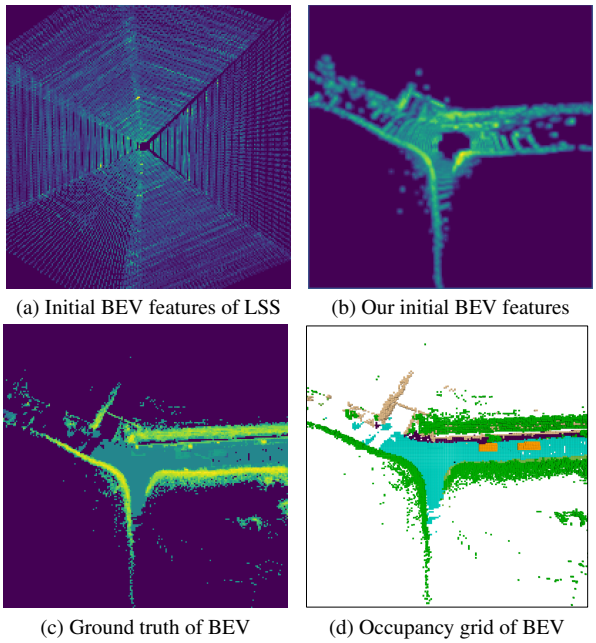


Figure 1. Comparisons of the mIoU and inference speed (FPS) of various 3D occupancy prediction methods on the Occ3D-nuScenes validation set. SDGOCC achieves higher accuracy and competitive inference speed.

Leveraging the complementary strengths of LiDAR and camera data is crucial for various 3D perception tasks. However, due to the heterogeneity between modalities, fusing LiDAR and camera data for 3D occupancy prediction remains challenging. Specifically, cameras provide rich semantic information but lack precise depth details, while LiDAR offers accurate depth information but only captures sparse data, potentially missing comprehensive scene details such as occluded objects. Existing methods often suffer from significant computational burdens (see Fig. 1), with some approaches attempting to leverage the LSS [26] pipeline for real-time performance. Although LSS simu-



(a) Initial BEV features of LSS (b) Our initial BEV features

(c) Ground truth of BEV (d) Occupancy grid of BEV

Figure 2. (a) BEV feature map of LSS with a shape of 200×200. We can observe that LSS has an extremely low utilization rate for BEV space. (b) The corresponding BEV features in SDG-OCC. Using depth and semantic information, the 2D-to-3D view transformation achieves efficient occupancy and utilization of the BEV features. (c) The corresponding BEV features in Ground Truth. (d) The corresponding BEV features in the occupancy grid.

lates the uncertainty of each pixel's depth through depth distribution (with depth intervals typically set to 0.5m), its sparse BEV representation allows only 50% of the grids to receive valid image features [15] (see Fig. 2 (a)). While increasing the depth interval can improve depth estimation accuracy to mitigate sparsity, it significantly increases computational demands.Additionally, while LiDAR can provide valuable geometric priors, fusion-based methods that process both point clouds and images simultaneously impose heavy computational burdens, thereby increasing the strain on real-time applications.

To address these issues, we propose a multimodal 3D semantic occupancy prediction framework, named SDG-OCC, which aims to achieve higher accuracy and competitive inference speed by fusing LiDAR information in the BEV perspective. In this framework, we introduce a semantic and depth-guided view transformation to replace normal BEV feature generation. Specifically, after extracting features from camera data and obtaining semantic segmentation masks and depth distributions through a multi-task head, we use the semantic masks and depth maps provided by LiDAR to construct virtual points via local diffusion and bilinear discretization. Combined with the depth distribution, these points are then projected into the BEV space. The comparison between LSS and our generated BEV features is shown in Fig. 2. The SDG view transformation significantly refines depth estimation accuracy and reduces redundant virtual point seeds, improving both the speed and accuracy of semantic occupancy.

Secondly, we introduce a fusion-to-occupancy-driven active distillation module. We first fuse LiDAR and camera features in the BEV space and then unidirectionally selectively transferred multimodal knowledge to image features based on LiDAR-identified regions. Our proposed SDG-Fusion, which includes only fusion, achieved SOTA performance on the Occ-3D-nuscenes [28] and SurroundOcc-nuScene [33] validation dataset. In comparison, SDG-KL, which combines fusion and unidirectional distillation, achieves real-time speed with a slight performance penalty.

Our contributions can be summarized as follows:

- We introduce a multimodal 3D semantic occupancy prediction framework, termed SDG-OCC, aimed at achieving higher accuracy and competitive inference speed by fusing LiDAR information in the BEV perspective.
- We propose a novel view transformation method that leverages the geometric and semantic information of point clouds to guide the 2D-3D view transformation. This significantly enhances the accuracy of depth estimation and improves both the speed and accuracy of semantic occupancy.
- We propose a fusion-to-occupancy-driven active distillation module that integrates multimodal features and selectively transfers multimodal knowledge to image features based on LiDAR-identified regions. Building on this, we present SDG-Fusion for high performance and SDG-KL for faster inference.
- Our method achieves SOTA performance with real-time processing on the occ3d-nuscenes dataset and shows comparable performance on the more challenging SurroundOcc-nuscenes validation dataset, demonstrating the effectiveness of our approach.

2. Related Work

2.1. Vision-Centric Occupancy Perception

Inspired by Tesla's autonomous driving perception system, vision-centric occupancy perception has garnered significant attention from both industry and academia. MonoScene [1] is a pioneering work that used only RGB inputs. TPVFormer [8] combines surround multi-camera inputs and uses transformer-based methods to lift features into a tri-perspective view space. SurroundOcc [33] extends high-dimensional BEV features into occupancy features and directly performs spatial cross-attention to generate geometric information. VoxFormer [12] introduces a two-stage transformer-based semantic scene completion framework, capable of outputting complete 3D volumetric semantics from 2D images alone. FlashOcc [40] trans-

forms the channel to height, lifting BEV output to 3D space, significantly improving operational efficiency. FBOcc [14] proposes a front-to-back view transformation module based on BEV features to address the limitations of different view transformations. Methods like UniOcc [25] and RenderOcc [24] use NeRF [32] to directly predict 3D semantic occupancy, but the rendering speed limits their efficiency. FastOcc [5] improves the occupancy prediction head to achieve a faster inference speed. COTR [20] builds compact 3D occupancy representations through explicit-implicit view transformation and coarse-to-fine semantic grouping. In this paper, we improve the speed and accuracy of 3D semantic occupancy prediction from the BEV space by incorporating the geometric and semantic information of the point cloud into the view transformation.

2.2. Multi-Modal Occupancy Perception

Multimodal occupancy perception leverages the strengths of multiple modalities to overcome the limitations of unimodal perception. OpenOccupancy[30] introduced a benchmark for LiDAR-camera semantic occupancy prediction. Inspired by BEVFusion, OccFusion [21] concatenates 3D feature volumes from different modalities along the feature channels, followed by convolutional layers to combine them. CO-Occ [23] introduced the Geometric and Semantic Fusion (GSFusion) module, identifying voxels containing both point cloud and visual information using k-nearest neighbors (KNN) search. OccGen [29] employs an adaptive fusion module to dynamically integrate occupancy representations from camera and LiDAR branches, using 3D convolutions to determine fusion weights for aggregating LiDAR and camera features. HyDR [34] proposes to integrate multimodal information in both perspective view (PV) and bird's-eye view (BEV) representation spaces. In this paper, we enhance view transformation by incorporating semantic segmentation masks and LiDAR depth maps to achieve higher occupancy accuracy. Additionally, we fuse BEV features from multimodal data and unidirectionally distill them into camera features, improving the accuracy and inference speed of 3D semantic occupancy prediction.

3. Methodology

3.1. Preliminary

Given joint input from multi-view images and LiDAR data, 3D occupancy prediction aims to estimate the occupancy state and semantic classification of 3D voxels surrounding the ego vehicle. Specifically, the input consists of a T-frame consequent sequence of images $X_C \in \mathbb{R}^{N_C \times H_C \times W_C \times 3}$ from N_C surround-view cameras and point clouds $X_L \in \mathbb{R}^{N_L \times (3+d)}$ as multimodal input, represented as $X=\{X_C, X_L\}$. Here H_C, W_C represent the height and width of the image, respectively, N_L denotes the number of point clouds and d denotes the initial additional features of the point cloud. Subsequently, we train a neural network to generate an occupancy voxel map $Y \in \mathbb{R}^{H \times W \times D \times C_N}$, where each voxel is assigned a label as unknown, occupied, or a semantic category from $\{C_0$ to $C_N\}$. Here, N denotes the total number of categories of interest, and H, W, D represent the volume dimensions of the entire scene.

3.2. Overall Architecture

An overview of SDGOCC is shown in Fig. 3. It mainly consists of four key modules: image feature encoder to extract image features, semantic and depth-guided view transformation to construct 2D-3D feature transformation, fusion-to-occupancy-driven active distillation for fusing multi-modal features and selectively transferring knowledge to the image features, and the occupancy prediction head for final output.

3.3. Image Encoder

The image feature encoder aims to capture multi-view features, providing a foundation for 2D-3D view transformation. Given RGB images from surround-view cameras, we first use a pre-trained image backbone network, such as classic ResNet [4] or strong Swin-Transformer [18], to extract multi-layer image features $F_C \in \mathbb{R}^{N_C \times C \times H \times W}$. These features are then aggregated using a feature pyramid network (FPN) [17], which combines fine-grained features and coarse-grained features and down-sampling them to a specific scale, typically 1/16.

3.4. SDG View Transformation

The way of converting image features to BEV features for better 3D perception was first proposed by the LSS pipeline. It constructs virtual points based on a predefined depth range for each pixel and predicts the depth distribution weight α and context feature c. Consequently, the feature representation of each pixel at depth d is given by $p_d = \alpha_d c$. Subsequently, all virtual points are projected into BEV space, where features within each column at height Z are aggregated to form the BEV features.

Although LSS can handle uncertainties and ambiguities in depth perception by using depth distributions to model pixel depth uncertainty, the resulting per-pixel feature remains large even with a 0.5 meter depth interval (e.g., approximately 0.4 million points per frame, which is an order of magnitude higher than point feature). Meanwhile, BEV features are highly sparse, with less than 50% of the image features being effective, which leads to suboptimal occupancy prediction performance. Reducing the depth interval can improve accuracy, but significantly increases computational burden and introduces irrelevant features, as most of the occupancy grid remains empty.

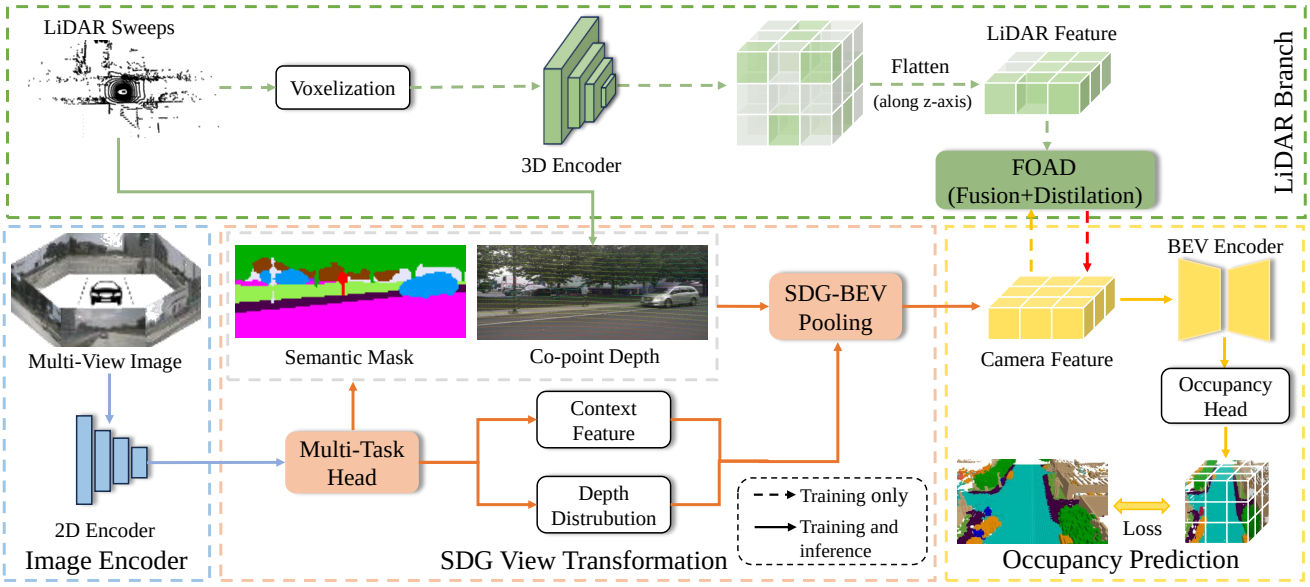


Figure 3. The overall architecture of SDG-OCC. T-frame multi-view images and corresponding point clouds are fed into the image and LiDAR backbones to extract features. Image features are processed by a multi-task head to generate semantic masks and depth distributions, combined with LiDAR depth maps to create virtual points for image BEV features. These features are fused with point cloud BEV features and selectively transfer multimodal knowledge to image features based on LiDAR-identified regions. Finally, enhanced features are processed by the occupancy prediction head to generate the occupancy map.

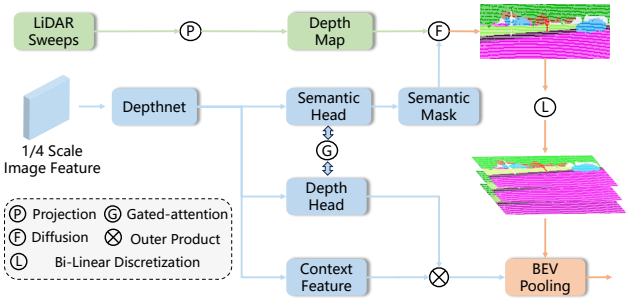


Figure 4. The overall architecture of the SDG view transformation. The image features processed by DepthNet are divided into texture features, depth features, and semantic features. Semantic features generate masks via a segmentation head, which are combined with depth maps from LiDAR for local diffusion and bilinear discretization to create virtual points. These points and their features receive pillar features from texture and depth features via the outer product, and then through BEV pooling to generate the final image BEV features.

To address this, we propose a novel view transformation that leverages the semantic information and sparse depth provided by LiDAR to guide the view transformation of features, as shown in Fig. 4. First, we extract features from multi-camera images and generate image semantic segmentation masks via a multi-task head, while simultaneously extracting image textual features and depth distribution weights, with the depth head and semantic head supplementing cross-task information through gated attention. To better utilize semantic information, we select 4x downsampled features for view transformation, as higher downsampling increases the semantic and depth ambiguity of pixels.

Given the differences in sparsity between images and point clouds, we combine image semantic segmentation masks and sparse projected depth maps provided by LiDAR to diffuse depth values within the same semantic category masks, generating a semi-dense extended depth map. This process is as follows:

$$D_{\text{temp}}(i,j) = \frac{\sum_{(p,q)\in N(i,j)} D(p,q)\cdot \mathbb{I}[M(p,q)=M(i,j)]}{\sum_{(p,q)\in N(i,j)} \mathbb{I}[M(p,q)=M(i,j)]} \tag{1}$$

where $N(i,j)$ represents the circle area with radius r around the current point, and $M(i,j)$ denotes the segmentation mask with N category. And $\mathbb{I}[M(p,q)=M(i,j)]$ checks if the semantic label at (p,q) matches that at (i,j), as follows:

$$\mathbb{I}[M(p,q)=M(i,j)] = \begin{cases} 1, & \text{if } M(p,q)=M(i,j)\neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and the final extended depth map $D_(i,j)$ replaces the original co-points of the $D_{\text{temp}}(i,j)$.

Due to the projection deviation from 2D pixels to 3D points, we apply bidirectional linear incremental discretization to the extended depth map to obtain discrete virtual points, enhancing the accuracy of depth estimation. These steps significantly reduce the number of virtual points,

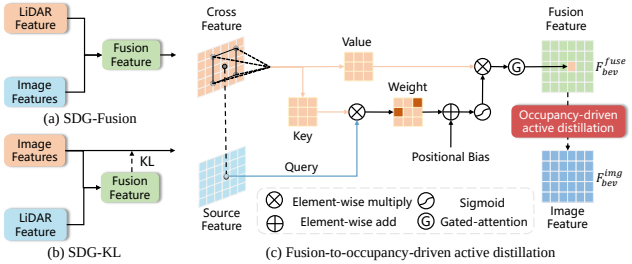


Figure 5. (a) The pipeline of SDG-Fusion. (b) The pipeline of SDG-KL. (c) The overall architecture of fusion-to-occupancy-driven active distillation. The source feature extracts neighborhood information from the corresponding pixels of the cross feature, where the source serves as the query, and the cross feature acts as both key and value for feature interaction. This interaction is dynamically refined through a gated attention mechanism to produce the fused feature. Subsequently, an occupancy-driven active distillation is used to unidirectionally integrate multimodal information into the image feature.

thereby improving inference speed. Finally, the image textual features F_t and depth distribution weights D_w are calculated by the outer product $F_t \otimes D_w$ to derive features for each virtual point and generate the BEV features F_{bev}^C of the camera through BEV pooling. This method effectively integrates semantic information with sparse depth data, significantly enhancing the accuracy of pixel depth estimation and improving the speed of view transformations.

3.5. Fusion-to-occupancy-driven Active Distillation

The LiDAR branch encompasses point cloud feature extraction, multimodal fusion, and occupancy-driven active distillation. Initially, the point cloud data undergoes voxelization and normalization to generate the initial features. We choose SPVCNN as our point-voxel feature encoder due to its efficiency in representing sparse point clouds while effectively preserving fine-grained details. Subsequently, we compress the voxel features at the corresponding scale to generate the LiDAR BEV features F_{bev}^L.

The complementary information from LiDAR and cameras is critical for 3D perception. A naive fusion method typically concatenates LiDAR and image BEV features along the channel dimension to enhance performance. However, feature misalignment due to extrinsic conflicts [2] limits the effectiveness of the fusion. Therefore, we propose a dynamic neighborhood feature fusion module. This module unidirectionally extracts neighborhood features from cross features and dynamically adjusts their weights into the source features using a gating attention mechanism.

External projection deviations during the BEV feature construction process for LiDAR and images result in misalignment between LiDAR and camera BEV features [2, 27]. Therefore, we adopt neighborhood attention from [3] to extract local patch features corresponding to the pixel from the cross features, and dynamically adjust the weights through gated attention to selectively enhance the fused feature representation. Specifically, the image features F_{bev}^C, as the source feature, are represented as a feature vector sequence $F_{img} \in \mathbb{R}^{n \times m}$, which is projected through a linear layer to obtain the query features $Q_s \in \mathbb{R}^{n \times q}$. Similarly, the LiDAR features F_{bev}^L as cross feature are projected to obtain the key $K_c \in \mathbb{R}^{n \times q}$ and value $V_c \in \mathbb{R}^{n \times v}$features. The local neighborhood features F_{neighbor} for a query point i are computed by the following equation:

$$F_{\text{neighbor}} = \sigma\left(\frac{Q_s^i \cdot (K_c^{n(i)})^T + B(i, n(i))}{\sqrt{v}}\right) \cdot V_c^i \quad (3)$$

where $n(i)$ represent the neighborhood with the size of k centered at the same position in the cross feature, $B(i, \rho(i))$ denotes the relative positional biases and σ denotes to Softmax. For each pixel in the feature map, we calculate the local neighborhood features. The fused features F_{bev}^{fuse} are then obtained from the local neighborhood features through gated attention, as follows:

$$F_{bev}^{fuse} = (\sigma(\text{Conv}(f_{\text{Avg}}(F_{\text{neighbor}}))) \cdot F_{\text{neighbor}} \quad (4)$$

where σ denotes the sigmoid function and Conv denotes linear transform matrix (e.g., 1x1 convolution), f_{Avg} refers Adaptive Average Pooling. The fused features F_{bev}^{fuse} are processed by the occupancy prediction head to obtain the SDG-Fusion model.

Additionally, to ensure real-time, we propose an occupancy-driven active distillation, where fused features are unidirectionally transferred to the image features. Specifically, LiDAR features are used as the source feature, while image features serve as the cross feature, resulting in LiDAR-dominant fused features, then we divide the space into two regions: the active region(AR), where both LiDAR and image features are occupied, and the inactive region(IR), where only LiDAR features are occupied. The details are as follows:

$$AR = (M_{\text{fused},i,j} = 1) \wedge (M_{\text{img},i,j} = 1)\,, \quad (5)$$

$$IR = (M_{\text{fused},i,j} = 1) \wedge (M_{\text{img},i,j} = 0)\,. \quad (6)$$

Where a value of 1 in $M_{\text{mode},i,j}$ indicates that the coordinate is occupied by the respective modality. Typically, the AR region is significantly larger than the IR region. To prevent the model from overemphasizing knowledge distillation in the AR region, we apply adaptive scaling based on the relative sizes of the AR and IR regions, as follows:

$$W_{I,i,j}^{(l_n)} = \begin{cases} \alpha, & \text{if } (i,j) \in AR, \\ \rho \times \beta, & \text{if } (i,j) \in IR, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where $\rho = \frac{N_{AR}}{N_{IR}}$ represents the relative importance of the IR over AR, α and β are the intrinsic balancing parameters, and N_{AR} and N_{IR} are the number of pixels in AR and IR, respectively.

The distillation loss between BEV feature from teacher F^t and student F^s are:

$$L_{\text{distill}} = \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} W_{i,j} \left(F_{bev}^{fuse} - F_{bev}^{C} \right)^2 \quad (8)$$

The network is trained with the sum of distillation and classification loss. The image features F_{bev}^{C} are processed by the occupancy prediction head to obtain the SDG-KL model.

3.6. Occupancy Prediction

To obtain 3D prediction outputs from coarse BEV features generated by view transformation, we propose an occupancy prediction system consisting of a BEV feature encoder and an occupancy prediction head. The BEV encoder uses several residual blocks for multi-scale feature diffusion and integrates a feature pyramid to acquire BEV features at the target scale. The occupancy prediction head extracts global features with multiple 3x3 convolutional layers and includes a channel-to-height transformation module. This module reshapes the BEV features from a $F_{out} \in \mathbb{R}^{B \times C \times H \times W}$ to $F_{final} \in \mathbb{R}^{B \times C_N \times D \times H \times W}$, where B, C, W, H, and D represent the batch size, channel number, class number, and the dimensions of the 3D space, respectively, with $C = C_N \times D$. Compared to traditional 3D encoders and occupancy prediction heads, this design significantly improves speed while maintaining comparable performance.

4. Experiments

We conduct experiments on the large-scale benchmark dataset Occ3D-nuScenes to validate the efficacy of our proposed methods. Additionally, we conduct ablation experiments to verify the effectiveness of each component in our method.

4.1. Datasets

Occ3D-nuScenes [28] is a large-scale autonomous driving dataset, which includes 1,000 urban traffic scenes under various conditions, the data is split into 700 training, 150 validation, and 150 testing scenes. The occupancy grid is defined within a range of -40m to 40m along the X and Y axes and -1m to 5.4m along the Z axis. The voxel size for occupancy labeling is set to 0.4m × 0.4m × 0.4m. The semantic labels include 17 categories consisting of 16 known object classes with an additional 'empty' class. Compared to Occ3D-nuScenes, SurroundOcc [33] is also based on the nuScenes dataset but its prediction range is from -50m to 50m for X and Y axes, and -5m to 3m along the Z axis, with the voxel label size of 0.2m × 0.2m × 0.2m.

4.2. Implementation Details

We use ResNet-50 as the default image backbone and SPVCNN as the LiDAR backbone. The model is trained on a GeForce RTX 4090 GPU using the AdamW optimizer with a learning rate of 1e-4 and gradient clipping. For semantic and depth-guided visual transformations, the bilinear incremental discretization range and the number of diffusion feature layers are set to 1 m and 8, respectively.

4.3. Comparing with SOTA methods

Occ3D-nuScenes. As shown in Table 1, we report the quantitative comparison of existing state-of-the-art methods for 3D occupancy prediction tasks on Occ3D-nuScenes. Most existing approaches are predominantly based on camera-only algorithms, with relatively few focusing on multi-sensor fusion. Our method, which employs a compact backbone and a lightweight LiDAR branch, achieves state-of-the-art performance in terms of mIoU and the majority of class-wise IoUs. Additionally, our approach achieves the best inference speed, meeting the real-time requirements of autonomous driving scenarios. The visualization on Occ3D-nuScenes validation set is shown in Fig.6. Compared to the baseline, our method effectively identifies categories that the baseline fails to correctly predict in both day and night scenarios. More visualizations can be found in supplementary material.

SurroundOcc-nuScenes. Tab. 2 provides a quantitative comparison on the SurroundOcc validation set, highlighting the performance of our method, SDG-OCC, compared to other approaches. Using both LiDAR and camera inputs, our method achieves SOTA performance on the SurroundOcc validation set, even with a finer voxel grid resolution. This success is attributed to our semantic and depth-guided view transformation, which enhances depth estimation accuracy and enables robust occupancy prediction across varying grid sizes. Notably, our method uses only a lightweight ResNet50 backbone and a lower resolution of 256×704, underscoring its effectiveness and efficiency.

Analysis of results within different ranges. We further evaluate different ranges surrounding the car to provide a comprehensive analysis. Fig. 7 clearly illustrates our mIoU and iou relative to the baseline FlashOcc. Short-range understanding is critical due to the limited reaction time for autonomous vehicles. Our method significantly outperforms the baseline in both mIoU and IoU. In long-range areas, where LiDAR data is sparse and few pixels define the depth of large regions, our approach still achieves superior IoU performance.

Method	Input	Backbone	Visible Mask	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation	Time(ms)
SurroundOcc [33]	C	R-101	✔	34.6	9.51	38.50	22.08	39.82	47.04	20.45	22.48	23.78	23.00	27.29	34.27	78.32	36.99	46.27	49.71	35.93	32.06	350.8
TPVFormer [8]	C	R-50	✔	34.2	7.68	44.01	17.66	40.88	46.98	15.06	20.54	24.69	24.66	24.26	29.28	79.27	40.65	48.49	49.44	32.63	29.82	320.0
OccFormer [42]	C	R-50	✔	37.4	9.15	45.84	18.20	42.80	50.27	24.00	20.80	22.86	20.98	31.94	38.13	80.13	38.24	50.83	54.3	46.41	40.15	-
VoxFormer [12]	C	R-101	✔	40.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FBOcc [15]	C	R-50	✔	42.1	14.30	49.71	30.0	46.62	51.54	29.3	29.13	29.35	30.48	34.97	39.36	83.07	47.16	55.62	59.88	44.89	39.58	-
PanoOcc [31]	C	R-101	-	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	370.4
FastOcc [5]	C	R-101	✔	40.75	12.86	46.58	29.93	46.07	54.09	23.74	31.10	30.68	28.52	33.08	39.69	83.33	44.65	53.90	55.46	42.61	36.50	221.2
BEVDet4D [7]	C	Swin-B	✔	42.5	12.37	50.15	26.97	51.86	54.65	28.38	28.96	29.02	28.28	37.05	42.52	82.55	43.15	54.87	58.33	48.78	43.79	1000.0
FlashOcc [40]	C	Swin-B	✔	43.52	13.31	51.62	28.07	50.91	55.69	27.46	31.05	29.98	29.20	38.86	43.68	83.87	45.63	56.33	59.01	50.63	44.56	909.1
COTR [20]	C	Swin-B	✔	46.2	**14.85**	53.25	**35.19**	50.83	57.25	35.36	34.06	33.54	**37.14**	38.99	44.97	84.46	**48.73**	57.60	61.08	51.61	46.72	699.3
HyDRa [34]	C+R	R-50	-	44.40	-	-	-	52.3	56.3	-	-	35.9	35.10	-	-	44.1	-	-	-	-	-	-	-
OCCFuion [22]	C+L	R-101	-	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	63.18	63.20	-
RadOcc-LC [41]	C+L	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	3333
SDG-KL	C+L	R-50	✔	50.16	12.26	57.12	23.69	58.77	62.74	34.55	36.19	50.1	32.05	49.89	51.24	84.1	46.05	57.2	61.45	69.56	65.78	**83**
SDG-Fusion	C+L	R-50	✔	**51.66**	13.21	57.77	24.3	**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**	133

Table 1. 3D Occupancy prediction performance on the Occ3D-nuScenes dataset. We present the IoU (geometry) and mean IoU (semantic) over categories and the IoUs (semantic) for different classes. The best scores for each class are highlighted in bold. In the Input, the C, L, and R denote camera, LiDAR, and radar, respectively. In the backbone, R represents ResNet, while Swin stands for Swin Transformer.

Method	Input	Backbone	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
SurroundOcc [33]	C	R-101	20.3	20.5	11.6	28.1	30.8	10.7	15.1	14.0	12.0	14.3	22.2	37.2	23.7	24.4	22.7	14.8	21.8
OccFormer [42]	C	R-101	20.1	21.1	11.3	28.2	30.3	10.6	15.7	14.4	11.2	14.0	22.6	37.3	22.4	24.9	23.5	15.2	21.1
C-CONet [30]	C	R-101	18.4	18.6	10.0	26.4	27.4	8.6	15.7	13.3	9.7	10.9	20.2	33.0	20.7	21.4	21.8	14.7	21.3
FB-Occ [15]	C	R-101	19.6	20.6	11.3	26.9	29.8	10.4	13.6	13.7	11.4	11.5	20.6	38.2	21.5	24.6	22.7	14.8	21.6
RenderOcc [41]	C	R-101	19.0	19.7	11.2	28.1	28.2	9.8	14.7	11.8	11.9	13.1	20.1	33.2	21.3	22.6	22.3	15.3	20.9
L-CONet [30]	L	-	17.7	19.2	4.0	15.1	26.9	6.2	3.8	6.8	6.0	14.1	13.1	39.7	19.1	24.0	23.9	25.1	35.7
FlashOcc* [40]	C	R-50	44.1	44.2	11.0	54.1	60.5	26.1	22.6	31.3	15.3	38.8	47.1	80.5	42.0	48.2	53.7	60.8	70.0
M-CONet [30]	C+L	R-101	24.7	24.8	13.0	31.6	34.8	14.6	18.0	20.0	14.7	20.0	26.6	39.2	22.8	26.1	26.0	26.0	37.1
Co-Occ [23]	C+L	R-101	27.1	28.1	16.1	34.0	37.2	17.0	21.6	20.8	15.9	21.9	28.7	42.3	25.4	29.1	28.6	28.2	38.0
OccFusion [22]	C+L+R	R-101	27.3	27.1	**19.6**	33.7	36.2	21.7	24.8	25.3	16.3	21.8	30.0	39.5	19.9	24.9	26.5	28.9	40.4
DAOcc [38]	C+L	R-50	30.5	30.8	19.5	34.0	38.8	25.0	**27.7**	29.9	**22.5**	23.2	31.6	41.0	25.9	29.4	29.9	35.2	43.5
SDG-KL	C+L	R-50	49.7	52.5	14.6	61.2	65.0	**34.9**	**32.0**	**48.5**	21.7	45.8	53.6	77.5	44.8	52.4	57.1	62.9	71.3
SDG-Fusion	C+L	R-50	**51.5**	**52.5**	11.4	**61.5**	**66.8**	34.5	32.0	41.7	21.6	**47.5**	**54.5**	**83.6**	**49.3**	**55.3**	**59.7**	**69.5**	**75.9**

Table 2. 3D Occupancy prediction performance on the SurroundOcc validation set. The best scores for each class are highlighted in bold. In the Input, the C, L, and R denote camera, LiDAR, and radar, respectively. * means the performance is achieved by our implementation using its official code.

Baseline	SDG	FOAD	iou(%)	mIoU(%)
✓			90.27	37.84
✓	✓		94.62	48.51
✓		✓	94.76	44.92
✓	✓	✓	95.35	51.66

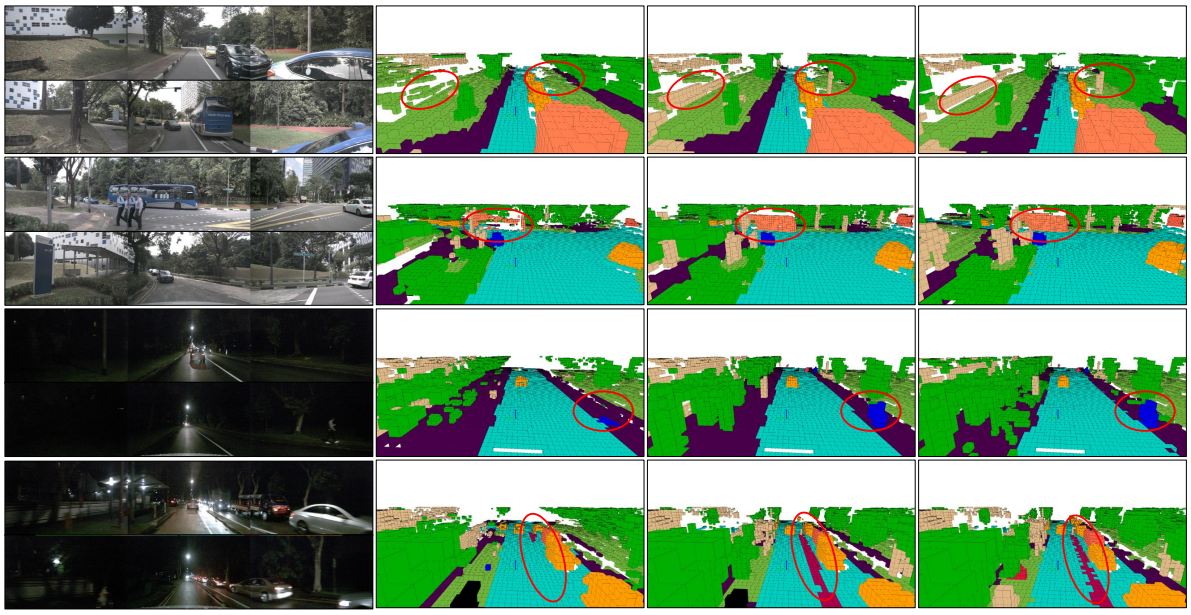
Table 3. Ablation study on the Occ3D-nuScenes dataset. SDG: Semantic and Depth-Guided View Transformations. FOAD: Fusion-To-Occupancy-Driven Active Distillation.

r	l	IoU(%)	mIoU (%)
1	4	95.34	51.15
1	8	95.35	**51.66**
1	12	95.38	51.28
2	4	95.35	50.4
2	8	**95.39**	51.03
2	12	95.30	51.12

Table 4. Ablation study of the hyperparameter used in SDG view transformation module on Occ3D-nuScenes.

4.4. Ablation study

The Effectiveness of Each Component. The results are shown in Table 3, we can observe that all components make their own performance contributions. The baseline achieves 90.27% of IoU and 37.84% of mIoU. We first integrated the Semantic and Depth-Guided (SDG) View Transformation into the baseline model, which brings 4.35% and 10.67% performance gain in IoU and mIoU. Fusion enhanced by integrating additional LiDAR information, the IoU and mIoU



Input Image Baseline Our Ground Truth

Figure 6. Qualitative results of SDG on the validation set of Occ3D-nuScenes. Each pair of rows displays results from day and low-light scene, respectively. Within each row, images from left to right represent the input images, baseline, our results, and the ground truth.

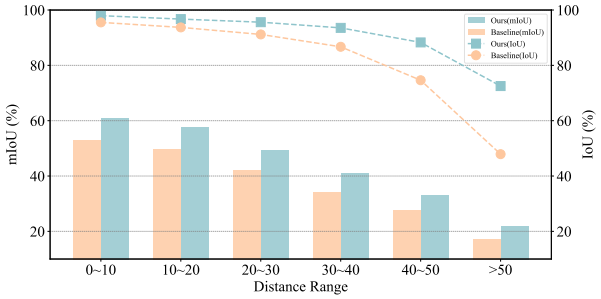


Figure 7. Distance-based evaluation on Occ3D-nuScenes. As the distance increases, the point cloud becomes sparse.

k	**3**	**5**	**7**	**9**
IoU (%)	95.32	**95.40**	95.35	95.26
mIoU (%)	51.22	51.40	**51.66**	51.08

Table 5. Ablation study of the hyperparameter used in feature fusion module on Occ3D-nuScenes. The K denotes the size of the pixel region corresponding to the neighborhood feature extraction.

have been significantly improved by 4.49% and 7.08%, respectively. By using both SDG and Fusion, outperforming the baseline by 5.19% of IoU and 13.82% of mIou.

The Effectiveness of SDG View Transformation. To further demonstrate the effect of SDG View Transformation, we conducted hyperparameter analysis experiments. In the SDG view transformation, the range r of bilinear growth discretization and the diffusion feature layers l control the virtual point generation of SDG. As shown in Table 4, lower depth precision (e.g., $r = 2$ and $l = 4$) results in slightly reduced performance compared to other configurations. However, excessive depth precision does not lead to additional gains, with the optimal performance observed at $r = 1$ and $l = 8$.

The Effectiveness of FOAD Module. We perform a hyperparameter analysis of the FOAD module. For neighborhood feature fusion, the parameter K controls the fusion of features from neighboring pixels. As shown in Table 5, increasing K does not consistently improve performance, with optimal results achieved at $K = 7$.

5. Conclusion

In this paper, we introduce a multimodal 3D semantic occupancy prediction framework, termed SDG-OCC, designed to achieve higher accuracy and competitive inference speed by fusing LiDAR information in the BEV perspective. To address the inaccurate depth estimation in view transformations, we propose a semantic and depth-guided view transformation method. This approach integrates pixel semantics and corresponding point depth through diffusion and bilinear discretization, effectively reducing invalid image features and significantly enhancing the speed and accuracy of semantic occupancy. Meanwhile, We propose a fusion-to-occupancy-driven active distillation that incorporates multimodal features and selectively transfers multimodal knowledge to image features based on LiDAR-identified regions. Our method achieves SOTA performance with real-time processing on the occ3d-nuscenes dataset and comparable performance on the more challenging SurroundOcc-nuscenes validation dataset, demonstrating its effectiveness.

References

[1] 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. 2

[2] Jiahui Fu, Chen Gao, Zitian Wang, Lirong Yang, Xiaofei Wang, Beipeng Mu, and Si Liu. Eliminating cross-modal conflicts in bev space for lidar-camera 3d object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16381–16387, 2024. 5

[3] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, 2023. 5

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. *arXiv preprint arXiv:2403.02710*, 2024. 3, 7

[6] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[7] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 7

[8] 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, 7

[9] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023. 1

[10] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17545–17555, 2023. 1

[11] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21694–21704, 2023. 1

[12] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 2, 7

[13] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: 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

[14] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 3

[15] 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, 7

[16] 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

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[19] 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. 1

[20] 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. 3, 7

[21] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: A straightforward and effective multi-sensor fusion framework for 3d occupancy prediction. *arXiv preprint arXiv:2403.01644*, 2024. 3

[22] 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. 7

[23] 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, 7

[24] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *arXiv preprint arXiv:2309.09502*, 2023. 3

[25] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai,

and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 3

[26] 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. 1

[27] Ziying Song, Lei Yang, Shaoqing Xu, Lin Liu, Dongyang Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. *arXiv preprint arXiv:2403.11848*, 2024. 5

[28] 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. 1, 2, 6

[29] 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*, 2024. 3

[30] 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. 1, 3, 7

[31] 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. 7

[32] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3

[33] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. c: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 2, 6, 7

[34] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. *arXiv preprint arXiv:2403.07746*, 2024. 3, 7

[35] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 1

[36] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 1

[37] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695. Springer, 2022. 1

[38] 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. 7

[39] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024. 1

[40] 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. 2, 7

[41] 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. 7

[42] 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. 7

[43] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021. 1