DAOcc: 3D Object Detection Assisted Multi-Sensor Fusion for 3D Occupancy Prediction

Zhen Yang* Yanpeng Dong Heng Wang
Beijing Mechanical Equipment Institute, Beijing, China

Abstract

Multi-sensor fusion significantly enhances the accuracy and robustness of 3D semantic occupancy prediction, which is crucial for autonomous driving and robotics. However, existing approaches depend on large image resolutions and complex networks to achieve top performance, hindering their application in practical scenarios. Additionally, most multi-sensor fusion approaches focus on improving fusion features while overlooking the exploration of supervision strategies for these features. To this end, we propose ***DAOcc****, a novel multi-sensor fusion occupancy network that leverages 3D object* ***d****etection supervision to* ***a****ssist in achieving superior performance, while using a deployment-friendly image feature extraction network and practical input image resolution. Furthermore, we introduce a BEV View Range Extension strategy to mitigate the adverse effects of reduced image resolution. As a result, our approach achieves new state-of-the-art results on the Occ3D-nuScenes and SurroundOcc datasets, using ResNet50 and a 256x704 input image resolution. Code will be made available at* https://github.com/AlphaPlusTT/DAOcc.

1. Introduction

3D semantic occupancy prediction (occ) is a critical task in autonomous driving [4, 47, 51] and robotic systems [14, 42, 44, 45], where accurately understanding the environment is essential for safe and efficient navigation. Reliable occupancy prediction requires not just accurate spatial data but also a comprehensive understanding of the environment's context. Achieving this necessitates the integration of data from multiple sensors. LiDAR provides precise 3D spatial information for obstacle detection, while cameras capture visual details like color and texture for a deeper understanding of the scene. By combining these complementary data sources, the accuracy and robustness of occupancy prediction are significantly enhanced.

*Corresponding author: yangzhen1324@163.com

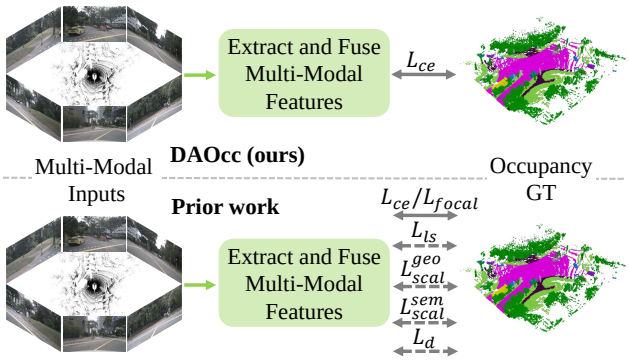

Figure 1. Unlike existing multi-modal methods [33, 34, 40, 46, 58], which rely on complex combinations of loss functions, including several or all of the focal loss $\mathcal{L}_{focal}$ [38], the scene-class affinity loss $\mathcal{L}^{geo}_{scal}$ and $\mathcal{L}^{sem}_{scal}$ [4], the lovasz-softmax loss $\mathcal{L}_{ls}$ [2], and the depth loss $\mathcal{L}_d$ [21], to achieve optimal performance, our approach only requires a single cross-entropy loss $\mathcal{L}_{ce}$ for the occupancy prediction task.

In existing works on multi-modal [33, 34, 40, 46, 49, 57, 58] or image-based [4, 12, 23, 39, 41, 47, 48, 50, 54] occ tasks, achieving superior performance often involves using extremely high image resolutions and complex image feature extraction networks (see Table 1 and Table 3), such as using 900×1600 resolution input image and ResNet101 [11] equipped with DCN [7, 62]. However, this approach significantly limits the deployment of top-performing occ models on edge devices due to the high computational demands. Compared to image, point cloud is much sparser. For example, in the training set of the nuScenes dataset, the maximum number of points in a single frame point cloud is only 34,880, which is equivalent to just 2.4% of the number of pixels in a 900×1600 resolution image. Therefore, how to effectively leverage point cloud data within a multi-modal occ framework remains to be further explored.

Moreover, we observe that most works [33, 40, 46, 49, 57, 58] on multi-modal occ primarily focuses on obtaining more effective fusion features, with insufficient exploration into the forms of supervision for these fused features. Although CO-Occ [34] introduces regularization based on im-

plicit volumetric rendering to supervise fused features, it only utilizes distance ground truth from the original point cloud data, failing to fully exploit the geometric and structural information inherent in point cloud features. In contrast, point cloud-based 3D object detectors [6, 31, 56] effectively leverage this information, achieving significantly better performance in 3D object detection tasks compared to image-based detectors [16, 26–28]. This observation suggests a new research direction: how to better exploit the unique strengths of point cloud data in multi-modal occ tasks.

Based on these observations, we propose **DAOcc**, a novel multi-modal occ framework that leverages 3D object **d**etection to **a**ssist in achieving superior performance while using a deployment-friendly image feature extraction network and practical input image resolution.

In constructing the baseline network structure of DAOcc, we adopt the most direct and straightforward approach: first, we extract features from images and point clouds, respectively. Given that depth estimation from monocular images is an ill-posed problem [13, 20] and that deformable attention modules are overly complex [10], we employ a simple method, similar to Harley *et al*. [9], to transform image features from 2D space to 3D volumetric space. Specifically, we project a set of predefined 3D points onto the 2D image feature plane and use bilinear interpolation to sample the corresponding 2D image features for these 3D points. Next, we adopt the same simplest fusion strategy as BEVFusion. [30], concatenating image and point cloud features, then performing a 2D convolution to fuse them, leading to a unified BEV features. Finally, we apply a fully convolutional BEV encoder with a residual structure to further fuse the unified BEV features, and subsequently employ the Channel-to-Height operation [54] to reshape the height of the unified BEV features.

To fully exploit the geometric and structural information inherent in point cloud features, we augment the baseline model's unified BEV features with 3D object detection supervision, thereby enhancing the discriminability of the unified BEV features. This renders the unified BEV features more sensitive to object boundaries and capable of perceiving relationships between internal object structures. Additionally, given the sparsity of point clouds, we extend the processing range of the point clouds and employ sparse convolutions [52] to mitigate the computational overhead introduced by this extension. We refer to this approach as BVRE (BEV View Range Extension). BVRE provides a larger BEV field of view, offering more contextual information and mitigating the adverse effects of reduced image resolution. It is worth noting that the 3D object detection supervision is only used as an auxiliary branch during training and can be removed during inference for the occ task.

As a result, the proposed DAOcc establishes the new state-of-the-art performance on the Occ3D-nuScenes and SurroundOcc benchmarks, while using ResNet50 and a 256×704 input image resolution. Specifically, on the Occ3D-nuScenes validation set, DAOcc achieves 53.82 mIoU when using the camera mask during training and 48.2 RayIoU without using the camera mask. Additionally, on the SurroundOcc validation set, DAOcc achieves 45.0 IoU and 30.5 mIoU.

In conclusion, our contributions are summarized as follows:

- We design a simple yet efficient multi-modal 3D semantic occupancy prediction baseline, eliminating the need for complex deformable attention modules as well as image depth estimation during feature fusion.
- We propose DAOcc, a novel multi-modal occupancy prediction framework that leverages 3D object detection to assist in achieving superior performance while using a deployment-friendly image feature extraction network and input image resolution.
- We introduce a BEV View Range Extension strategy, which provides a larger BEV field of view, offering more contextual information and mitigating the adverse effects of reduced image resolution.
- We establish the new state-of-the-art performance on Occ3D-nuScenes dataset and SurroundOcc dataset, while using ResNet50 and a 256×704 input image resolution.

2. Related Work

3D Occupancy Prediction: 3D occupancy prediction aims to map all occupied voxels in the environment and assign semantic labels, thus providing more fine-grained perception results.

Camera-based 3D occupancy prediction has gained significant attention due to its cost-effectiveness. MonoScene [4] is the first work to infer the dense occupancy and semantics of outdoor and indoor scenes from a single RGB image. Since full perception of the surrounding environment is crucial for making accurate decisions in autonomous driving and robotics, most of the recent work is based on surround view image input. TPVFormer [15] proposes an efficient tri-perspective view (TPV) representation that combines BEV with two additional perpendicular planes to provide a 3D perception result with multi-view image inputs. Given the insufficient fine-grained semantic information of the TPV representation, OccFormer [59] leverages dense 3D features and proposes a dual-path transformer-based occupancy network. However, both TPVFormer and OccFormer use sparse LiDAR points as supervision, resulting in sparse occupancy predictions. To obtain dense occupancy prediction, OpenOccupancy [46], SurroundOcc [48] and Occ3D [41] developed methods for generating dense occupancy labels and established benchmarks on their respective proposed datasets. It is

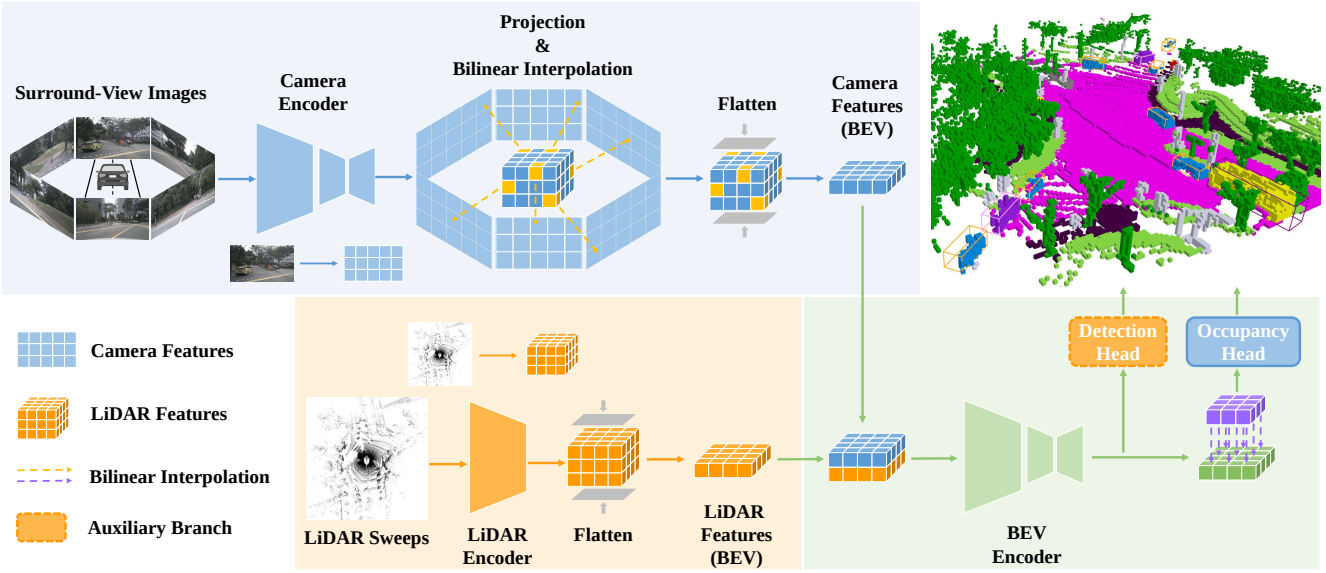

Figure 2. **Overview of our proposed DAOcc.** We first introduce the BVRE strategy to enrich the spatial contextual semantic information by broadening the perceptual range from the BEV perspective. Then feature extraction is performed through the multi-modal occupancy prediction network illustrated in the above figure. Furthermore, to fully leverage the inherent geometric structures within point cloud features, we incorporate 3D object detection as auxiliary supervision.

worth noting that SurroundOcc does not introduce manual annotation. The emergence of these benchmarks further promoted the development of 3D occupancy prediction. PanoOcc [47] adopts voxel query and a coarse-to-fine approach to learn a unified occupancy representation. FlashOcc [54] uses BEV features with 2D convolution to achieve efficient feature extraction and proposes the channel-to-height transformation to lift the output of BEV to 3D space. SparseOcc [29] leverages a sparse voxel decoder to reconstruct the sparse geometry of a scene and uses sparse queries to predict masks and labels. OSP [39] treats the scene as a collection of points and uses flexible sampling to allow the model to pay special attention to certain areas. However, all camera-based methods inevitably need to cope with the challenges of adverse lighting and weather conditions, and many works inevitably lack robustness due to the ill-posed monocular depth estimation problem [13, 20].

While camera-based occupancy prediction has demonstrated promising results, multi-modal approaches offer greater reliability and robustness, making them indispensable for practical applications in autonomous driving and robotics. Due to the fact that cameras are susceptible to changing lighting and weather conditions, OccFusion [33] enhances the accuracy and robustness of the occupancy network by integrating features from LiDAR and radar. In concurrent work, OccFusion [58] projects the preprocessed denser and more uniform point cloud onto the image plane to establish a mapping relationship, and performs deformable attention [63] to fuse the corresponding features. Although OccFusion [58] avoids depth estimation, using deformable attention incurs a greater computational burden. Hydra [49] extends FB-BEV [24] to the camera-radar fusion domain and improves the accuracy of depth estimation with the help of radar features. Co-Occ [34] employs K-nearest neighbor search in selecting neighboring camera features to enhance the corresponding LiDAR features and proposes a regularization based on implicit volume rendering. However, this regularization only utilizes the distance ground truth of the point cloud and does not leverage its intrinsic geometry information. EFFOcc [40] proposes an efficient and lightweight multi-modal 3D occupancy network, but it significantly relies on the 3D detection pretraining of the LiDAR branch for optimal results. This dependence, however, limits the flexibility of the network structure design. Specifically, to maximize the benefits of the 3D detection pretraining, EFfOcc has to adjust its network architecture to align more closely with well-established LiDAR detection networks. In addition, the inherently ill-posed nature of the monocular depth estimation problem inevitably renders the image branch of EFFOcc less robust [13]. In this work, we introduce a simple yet efficient multi-modal occupancy prediction network, eliminating the need for complex deformable attention [22, 63] as well as image depth estimation during feature fusion, and does not require 3D object detection pretraining. Furthermore, to fully exploit the inherent geometry information in point cloud features, we add supervision of 3D object detection as an auxiliary

branch on the fused features.

Multi-Modal 3D Object Detection: Recent multi-modal 3D object detection methods [5, 18, 19, 30, 43] mainly focus on learning effective BEV feature representations. TransFusion [1] proposes a two-stage transformer-decoder based detection head and applies cross attention to obtain image features for each object query. BEVFusion [30] proposes an efficient and generic multi-task multi-sensor fusion framework which unifies multi-modal features in the shared bird's-eye view (BEV) representation space, and introduces a specialized kernel to speed up the BEV pooling operation. In concurrent work, to improve the robustness of the LiDAR-camera fusion framework to sensor failures, such as missing LiDAR sensor input, BEVFusion decomposes the LiDAR-camera fusion into two streams that can independently output perception results, and performs feature fusion after the two streams. DAL [13] follows the concept of 'Detecting As Labeling' and decouples the fused features during classification and regression. Specifically, it employs the fused features for classification, whereas it utilizes the point cloud features exclusively for regression. In this work, we incorporate the simple yet effective feature fusion approach of BEVFusion and introduce 3D object detection as an auxiliary branch during training.

3. Proposed Method

3.1. Overall Framework

Our goal is to fully exploit the advantages of point cloud features for multi-modal occupancy prediction. Previous multi-modal studies have not fully exploited this and can only achieve superior performance through more complex image feature extraction networks and larger input image resolutions. The overall framework of our proposed DAOcc is illustrated in Figure 2. DAOcc takes surround images and their corresponding time-synchronized point cloud as input, and obtains the features of the image and point cloud through the Camera Encoder and LiDAR Encoder, respectively. The 2D image features are transformed into the 3D voxel space by projection and sampling. Subsequently, the image and point cloud features in the 3D space are compressed along the height dimension to generate the corresponding BEV features. A simple 2D convolution is then applied for feature fusion, and the fully convolutional BEV Encoder encodes these fused features to obtain the final BEV representations. Finally, the occupancy head uses the Channel-to-Height operation [54] to restore the height of the BEV representations, resulting in final 3D voxel space representations that can be utilized for occupancy prediction. These modules collectively form the basic network architecture of DAOcc, which will be elaborated in Section 3.2.

On the foundation of the basic network, we introduce the BVRE strategy (see Section 3.3) to compensate for the information loss that arises from the reduction in image resolution. This strategy aims to enrich the spatial contextual semantic information by broadening the perceptual range from the Bird's Eye View (BEV) perspective. Furthermore, to fully leverage the inherent geometric structures within point cloud features, we incorporate 3D object detection as auxiliary supervision (see Section 3.4). This auxiliary supervision not only enhances the discriminability of the fused features but also leads to a very concise overall training loss (see Section 3.5) for our proposed framework.

3.2. Basic Network

LiDAR Encoder. The method of embedding raw LiDAR points into 3D voxelized features is consistent with BEVFusion [30]. We first voxelize the point cloud, retaining a maximum of 10 points per voxel, resulting in a 3D voxel grid of size $D \times H \times W$.The feature representation for each voxel is obtained by averaging the features of all points within it. Next, we apply 3D sparse convolutions [52] to encode these voxel features, generating spatially compressed LiDAR voxel features $F_l \in \mathbb{R}^{C \times \frac{D}{16} \times \frac{H}{8} \times \frac{W}{8}}$, where C represents the feature dimensions.

Camera Encoder. For image feature extraction, taking surround images as input, we first use ResNet50 [11] as the backbone to extract multi-scale features, denoted as $F_{ms} = \{F_{ms \times \frac{1}{8}}, F_{\frac{1}{16}}, F_{\frac{1}{32}}\}$, where $F_{\frac{1}{x}}$ represent features extracted after $x\times$ downsampling. Then, we employ the Feature Pyramid Network (FPN) [25] as the neck to aggregate these multi-scale features. The output feature map F_{c_p} has a shape of $N_p \times C_p \times \frac{H_p}{8} \times \frac{W_p}{8}$, where H_p and W_p represent the input resolution of the image, and C_p and N_p indicate the number of channels and the number of surround images, respectively.

Projection and Interpolation. For image-related occupancy prediction, a key step is to transform image features from 2D image plane to 3D volume space. Most existing methods use monocular depth estimation [12, 40, 46, 54] or deformable attention [41, 47, 48, 58]. However, monocular depth estimation is inherently an ill-posed problem [13, 20], while deformable attention imposes a significant computational burden [10]. Given these limitations, we use a simple yet effective projection and sample approach similar to Harley *et al*. [9]. Concretely, we first predefine a 3D voxel grid with a shape of $Z \times \frac{H}{8} \times \frac{W}{8}$, where Z represents the number of voxels along the z-axis. The center point of each voxel is then projected onto the image feature plane, and only points that fall within both the image feature plane and the camera's field of view are retained. Next, the sub-pixel projection positions of the retained points are bilinearly interpolated to generate the image features corresponding to each voxel. For voxels located in the overlapping viewing regions of the surround cameras, we average the image fea-

tures from the two corresponding cameras to obtain the final feature for each voxel. The output camera voxel features can be denoted as $F_c \in \mathbb{R}^{C \times Z \times \frac{H}{8} \times \frac{W}{8}}$.

BEV Encoder. Given the fused feature F_f, we further refine F_f by passing it through three blocks of ResNet18 [11], resulting in two feature maps, F_{f0} and F_{f2}, extracted at two scales from the first and last blocks respectively. Then, similar to FPN [25], we apply bilinear upsampling to F_{f2} and concatenate it with F_{f0} along the feature dimension. Finally, we fuse the features of different scales using a convolution block. The output refined BEV features can be denoted as $F_r \in \mathbb{R}^{C_r \times \frac{H}{8} \times \frac{W}{8}}$.

3.3. BVRE

Considering the sparsity of point clouds, the increase in computational cost resulting from expanding the processing range of point clouds is very small compared to images. Therefore, we extend the point cloud range to provide more 3D spatial context, compensating for the information loss caused by lower image resolution. However, arbitrarily setting the XY range of the point cloud or voxel resolution may result in misalignment between the BEV features and the occupancy ground truth. Assuming that the voxel resolution of the occupancy grid is res_o, the expanded point cloud data in BEV covers a rectangular region with x-coordinates ranging from $-x$ to x and y-coordinates ranging from $-y$ to y, and has a voxel resolution of res_p. Thus, x and y must be integer multiples of res_o, and res_o must also be an integer multiple of res_p. Otherwise, the spatial resolution along the z-axis for each feature vector in the BEV representation will not be equal to res_o. To avoid complex manual design, we adopted coordinate transformation and interpolation, as illustrated in the purple part of Figure 2, which can be formulated as follows:

$$F_{occ} = \text{GridSample}(F_r, \text{Norm}(T_{o2p} \times P_o)) \quad (1)$$

where P_o is a set of predefined points in the XY plane of the coordinate system of the occupancy annotation, each point is located at the center of an occupancy grid in the XY plane. T_{o2p} is the transformation matrix from occupancy coordinates to Lidar coordinates. The function of Norm is to scale the coordinate values to a range from -1 to 1.

3.4. Detachable Auxiliary Head

To fully exploit the inherent geometry information in point clouds, we add a separable auxiliary detection network to the fused features, which will enable the features to benefit more from the geometry information. Meanwhile, the auxiliary detection task is related to the occupancy prediction task, thus providing multiple regularization effects during optimization. For simplicity, we use the one-stage CenterHead introduced in CenterPoint [53] as the auxiliary detection head. Given the refined BEV feature F_r, we utilize two convolutional layers to generate a keypoint heatmap $HM = p_{xy}$ from the BEV perspective, and apply a Gaussian kernel [17, 60] to map the center points of all ground truth 3D boxes onto a target heatmap T. The training objective is formulated as a focal loss [38] based on Gaussian heatmaps:

$$\mathcal{L}_{cls} = \frac{-1}{N} \sum_{ij} \begin{cases} (1-p_{ij})^{\alpha} \log(p_{ij}) & \text{if } y_{ij} = 1 \\ (1-y_{ij})^{\beta} (p_{ij})^{\alpha} \log(1-p_{ij}) & \text{otherwise} \end{cases} \quad (2)$$

where p_{ij} and y_{ij} represent the predicted score and ground truth of the heatmap at location (i, j), respectively. N is the number of objects in a point cloud, and α and β are hyperparameters of the focal loss [38]. Similarly, for the offset of the center point, as well as the height-above-ground, 3D dimensions and yaw of the 3D bounding box, we use separate convolutional layers to predict each of these parameters, and then apply the L1 loss for supervision:

$$\mathcal{L}_{loc} = \frac{1}{N} \sum_{k=1}^{N} |\hat{s}_k - s_k| \quad (3)$$

where $\hat{s}_k$ and s_k represent the predicted value and the ground truth respectively. The above two tasks enable the network to perceive object boundaries more precisely, which in turn facilitates achieving more precise occupancy predictions. The total loss for the auxiliary detection head can be denoted as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{loc} \quad (4)$$

3.5. Overall Objective Function

Most existing methods rely on complex combinations of loss functions [23, 33, 34, 39, 40, 46, 47, 50, 55, 58], such as a combination of several or all of the focal loss [38], the scene-class affinity loss [4], the dice loss, the lovasz-softmax loss [2], the depth loss [21], etc., to achieve the expected performance. In contrast, our approach, utilizing the proposed auxiliary supervision for 3D object detection, requires only a simple cross-entropy loss $\mathcal{L}_{ce}$ for occupancy prediction. The total loss for our framework can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{det} \quad (5)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Occ3D-nuScenes [41] is a large-scale benchmark for 3D occupancy prediction, containing 700 scenes for training and 150 for validation, each lasting 20 seconds with annotations at 2 Hz. The dataset covers a perception range of [-40m, 40m] in the X and Y directions and [-1m, 5.4m] along the z-axis, which is discretized into voxels of [0.4m,

Method	mIoU	Modality	Input Size	2D Backbone	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
CTF-Occ [41]	28.53	C	928 × 1600	R101-DCN	8.1	39.3	20.6	38.3	42.2	16.9	24.5	22.7	21.1	23.0	31.1	53.3	33.8	38.0	33.2	20.8	18.0
BEVDetOcc(2f) [12]	42.02	C	512 × 1408	Swin-B	12.2	49.6	25.1	52.0	54.5	27.9	28.0	28.9	27.2	36.4	42.2	82.3	43.3	54.6	57.9	48.6	43.6
PanoOcc(4f) [47]	42.13	C	864 × 1600	R101-DCN	11.7	50.5	29.6	49.4	55.5	23.3	33.3	30.6	31.0	34.4	42.6	83.3	44.2	54.4	56.0	45.9	40.4
FB-OCC(16f) [23]	48.90	C	960 × 1760	VoVNet-99	14.3	57.0	38.3	57.7	62.1	34.4	39.4	38.8	39.4	42.9	50.0	86.0	50.2	60.1	62.5	52.4	45.7
FlashOcc(2f)* [54]	43.52	C	512 × 1408	Swin-B	13.4	51.1	27.7	51.6	56.2	27.3	30.0	29.9	29.8	37.8	43.5	83.8	46.6	56.2	59.6	50.8	44.7
OSP [39]	39.41	C	900 × 1600	R101-DCN	11.2	47.3	27.1	47.6	53.7	23.2	29.4	29.7	28.4	32.4	39.9	79.4	41.4	50.3	53.2	40.5	35.4
DHD(2f) [50]	45.53	C	512 × 1408	Swin-B	14.1	53.1	32.4	52.4	57.4	30.8	35.2	33.0	33.4	37.9	45.3	84.6	48.0	57.4	60.3	52.3	46.2
RadOcc [57]	49.38	C+L	512 × 1408	Swin-B	10.9	58.2	25.0	57.9	62.9	34.0	33.5	50.1	32.1	48.9	52.1	82.9	42.7	55.3	58.3	68.6	66.0
OccFusion [33]	46.67	C+L+R	900 × 1600	R101-DCN	12.4	50.3	31.5	57.6	58.8	34.0	41.0	47.2	29.7	42.0	48.0	78.4	35.7	47.3	52.7	63.5	63.3
OccFusion [58]	48.74	C+L	900 × 1600	R101	12.4	51.8	33.0	54.6	57.7	34.0	43.0	48.4	35.5	41.2	48.6	83.0	44.7	57.1	60.0	62.5	61.3
HyDRa [49]	44.4	C+R	256 × 704	R50	-	-	-	52.3	56.3	-	35.9	35.1	-	-	44.1	-	-	-	-	-	-
EFFOcc [40]	52.62	C+L	512 × 1408	Swin-B	14.7	58.6	34.7	60.5	65.5	36.0	40.4	51.5	41.0	48.5	54.6	84.2	46.4	57.9	60.8	70.9	68.4
DAOcc	**53.82**	C+L	**256 × 704**	**R50**	12.4	59.6	38.4	61.9	67.1	35.3	48.2	59.1	43.5	50.9	56.3	83.0	44.7	56.7	59.9	70.0	68.1

Table 1. **3D Semantic occupancy prediction performance on Occ3D-nuScenes validation set.** (xf) means use x frames for temporal fusion. * denotes the usage of exponantial moving average (EMA). The C, L, and R denote camera, LiDAR, and radar, respectively.

Method	RayIoU	Modality	Input Size	2D Backbone	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	mIoU
BEVFormer(4f) [22]	32.4	C	900 × 1600	R101	26.1	32.9	38.0	39.2
RenderOcc [35]	19.5	C	512 × 1408	Swin-B	13.4	19.6	25.5	24.4
SimpleOcc [8]	22.5	C	336 × 672	R101	17.0	22.7	27.9	31.8
BEVDetOcc(8f) [12]	32.6	C	384 × 704	R50	26.6	33.1	38.2	39.3
FB-Occ(16f) [23]	33.5	C	256 × 704	R50	26.7	34.1	39.7	39.1
SparseOcc(16f) [29]	36.1	C	256 × 704	R50	30.2	36.8	41.2	30.9
Panoptic-FlashOcc [55]	35.2	C	256 × 704	R50	29.4	36.0	40.1	29.4
Panoptic-FlashOcc(8f) [55]	38.5	C	256 × 704	R50	32.8	39.3	43.4	31.6
DAOcc	**48.2**	C+L	256 × 704	R50	44.4	48.7	51.6	**47.88**

Table 2. **3D Semantic occupancy prediction performance on Occ3D-nuScenes validation set without camera mask.** (xf) means use x frames for temporal fusion. The C and L denote camera, LiDAR, respectively.

Method	IoU	mIoU	Modality	Input Size	2D Backbone	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [4]	24.0	7.3	C	900× 1600	R101-DCN	4.0	0.4	8.0	8.0	2.9	0.3	1.2	0.7	4.0	4.4	27.7	5.2	15.1	11.3	9.0	14.9
BEVFormer [22]	30.5	16.7	C	900× 1600	R101-DCN	14.2	6.5	23.4	28.2	8.6	10.7	6.4	4.0	11.2	17.7	37.2	18.0	22.8	22.1	13.8	22.2
SurroundOcc [48]	31.4	20.3	C	900× 1600	R101-DCN	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 [59]	29.9	20.1	C	896× 1600	R101	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 [46]	26.1	18.4	C	896× 1600	R101	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 [23]	31.5	19.6	C	896× 1600	R101	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 [35]	29.2	19.0	C	896× 1600	R101	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
LMSCNet [37]	36.6	14.9	L	-	-	13.1	4.5	14.7	22.1	12.6	4.2	7.2	7.1	12.2	11.5	26.3	14.3	21.1	15.2	18.5	34.2
L-CONet [46]	39.4	17.7	L	-	-	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
M-CONet [46]	39.2	24.7	C+L	896× 1600	R101	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 [34]	41.1	27.1	C+L	896× 1600	R101	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 [33]	44.7	27.3	C+L+R	900× 1600	R101-DCN	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	**45.0**	**30.5**	C+L	**256× 704**	**R50**	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

Table 3. **3D Semantic occupancy prediction performance on SurroundOcc validation set.** The C, L, and R denote camera, LiDAR, and radar, respectively.

0.4m, 0.4m] in size. Each occupied voxel is assigned one of 17 semantic labels, including 16 common classes and a general object class 'others'. The data includes LiDAR point clouds and RGB images from six surround cameras, enabling dense voxel-wise annotation for 3D scene understanding. In addition, the dataset also provides visibility

masks for LiDAR and camera modes, and it can be used for training.

SurroundOcc [48] is an automatically generated occupancy prediction dataset without human annotation. It is built upon nuScenes [3] and generates dense occupancy ground truth utilizing existing 3D detection and 3D semantic segmentation labels. The dataset employs the training set to train the model and utilizes the validation set for evaluation purposes. The occupancy prediction range is set to [-50m, 50m] for the X and Y axes, and [-5m, 3m] for the z-axis. The final output occupancy grid has a shape of 200×200×16, with a voxel size of [0.5m, 0.5m, 0.5m].

Evaluation metrics: Following the previous methods, when training Occ3D-nuScenes with the camera visible mask, we use the mean Intersection over Union (mIoU) for fair comparison with previous methods. When the camera visible mask is not used during training, we report an additional semantic segmentation metric, RayIoU [29], which was recently proposed. For SurroundOcc, we report semantic segmentation performance using IoU and mIoU, where IoU is the intersection over union of occupied voxels. This IoU is used as an evaluation metric for the scene completion (SC) task by ignoring the semantic categories of occupied voxels.

4.2. Implementation Details

We implement our method using PyTorch [36] and train all models on 8 NVIDIA RTX 4090 GPUs. We utilize ResNet-50 [11], which is pre-trained on nuImages [3], as the image backbone and crop the input image size to 256×704. In the LiDAR branch, we voxelize 10 LiDAR scans and start training from scratch. We employ the AdamW [32] optimizer with a cosine annealing learning rate scheduler, including a warmup phase, and set an initial learning rate of 2e-4. We apply data augmentation to both the input image and the ground truth 3D bounding boxes, and adopt CBGS [61].

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

Comparison on Occ3D-nuScenes with camera mask. As presented in Table 1, we compare our DAOcc with both single-modal and multi-modal methods on Occ3D-nuScenes, all results are either implemented directly by their authors or reported based on their official code implementations. All methods use the camera visible mask during training, except for CTF-Occ [41]. Our DAOcc establishes a new state-of-the-art performance, achieving an outstanding 53.82 mIoU, while utilizing a 2D backbone ResNet50 and 256×704 input image resolution. Compared with the current state-of-the-art multi-modal method EFFOcc [40], our DAOcc outperforms EFFOcc by 1.2 mIoU. Moreover, while EFFOcc employs Swin-B and an input image resolution of 512×1408, DAOcc utilizes only ResNet50 and an input image resolution of 256×704. In addition, EFFOcc requires the LiDAR branch to perform 3D object detection pre-training, which results in an improvement of approximately 3 mIoU, in order to achieve optimal results, whereas DAOcc does not require 3D detection pretraining.

	BEV View Range	Voxel Resolution (BEV)	mIoU
(a)	$[-41.4, 41.4]$	0.075×0.075	50.76
(b)	$[-51.2, 51.2]$	0.1×0.1	49.98
(c)	$[-54.0, 54.0]$	0.075×0.075	**51.26**

Table 4. **Performance comparison between different BEV processing ranges and voxel sizes.**

Comparison on Occ3D-nuScenes without camera mask. To the best of our knowledge, no existing multi-modal occupancy prediction work has used RayIoU metrics to report model performance, so we mainly compare with image-based methods. As shown in Table 2, thanks to the multi-modal input and the effectiveness of our proposed method, our DAOcc is significantly ahead of image-based methods, although many of these image-based methods use multi-frame temporal fusion. It should be noted that some image-based methods based on monocular depth estimation, such as Panoptic-FlashOcc [55], only require images as input during the inference phase, but require additional point cloud data input to provide depth information supervision during the training phase, so these methods can also be regarded as multi-modal during the training phase.

Comparison on SurroundOcc. Table 3 presents a quantitative comparison on SurroundOcc validation set, showcasing the performance of our DAOcc against other methods. By leveraging both LiDAR and camera inputs, DAOcc achieves a remarkable 3.2 mIoU increase over OccFusion [33], which not only employs camera and LiDAR but also incorporates radars. Notably, OccFusion utilizes R101-DCN with a high-resolution input of 900×1600, whereas our method employs a more lightweight Resnet50 backbone and a lower resolution of 256×704. This underscores the effectiveness and efficiency of our proposed method.

4.4. Ablation Studies

Ablation studies are conducted on Occ3D-nuScenes using the camera mask. Unless otherwise specified, all experiments presented in this section are trained for 15 epochs only.

We propose the BVRE strategy to enrich the spatial contextual semantic information by broadening the perceptual range from the BEV perspective. As shown in Table 4, comparing (a) and (c), increasing the BEV view range from [-41.4, 41.4] to [-54.0, 54.0] can bring a performance improvement of 0.5 mIoU. In addition, by comparing (a) and (b), we can find that the size of the voxel has a great impact

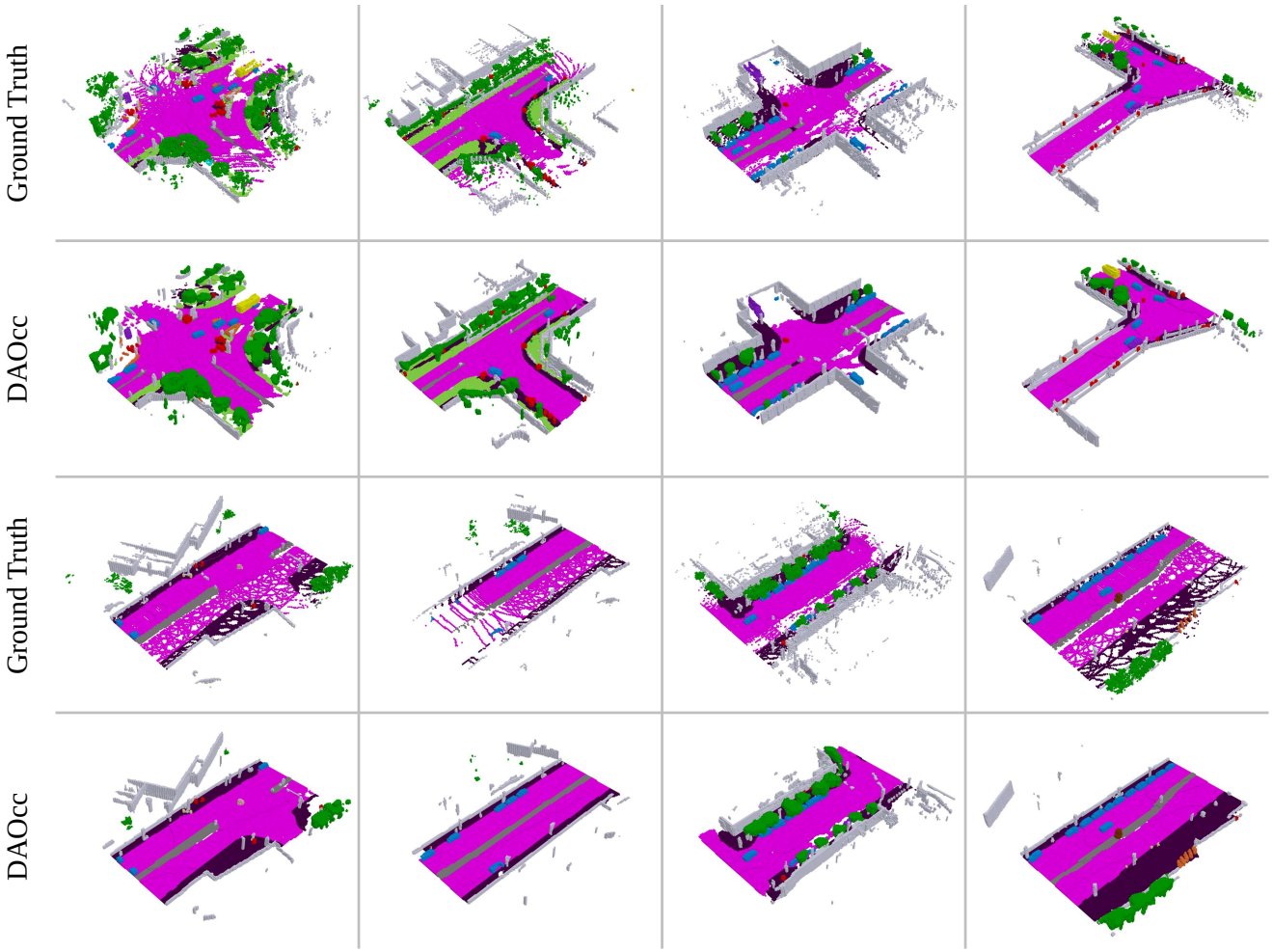

Figure 3. Qualitative visualizations on Occ3D-nuScenes validation set without camera mask.

Detection Auxiliary	mIoU	bicycle	motorcycle	traffic cone
	51.26	20.60	38.89	36.65
✓	**52.82**	**36.35**	**45.48**	**41.98**

Table 5. **Ablation study on the auxiliary detection head.**

Detection Auxiliary	EP15	EP24	CBGS (EP6)	mIoU
✓	✓			52.82
✓		✓		53.39
✓			✓	**53.82**

Table 6. **Ablation for obtaining the final performance.**

on the model's performance. Setting a smaller voxel size will bring obvious performance improvements but will also impose a greater computational burden.

To fully leverage the inherent geometric structures within point cloud features, we incorporate 3D object detection as auxiliary supervision. As shown in Table 5, by adding auxiliary supervision for 3D object detection, we can achieve an improvement of 1.56 mIoU, demonstrating the effectiveness of our method. Given that the gain in object detection for foreground objects will be greater, we present the three categories with the highest gain in foreground objects, which are the bicycle (+15.75 mIoU), the motorcycle (+6.59 mIoU) and the traffic cone (+5.33 mIoU). This further proves the rationality of the performance improvement of our method.

Table 6 provides ablation experiments to obtain the final performance. First, by increasing the training time from 15 epochs to 24 epochs, an improvement of 0.57 mIoU can be achieved. Then, by using CBGS [61], another improvement

of 0.43 mIoU can be achieved. When using CBGS, since resampling increases the number of samples in each epoch, for the sake of a fair comparison, we limited the training to 6 epochs to approximate the total training time of 24 epochs without CBGS.

4.5. Visualization

Figure 3 presents the visualizations of DAOcc on Occ3D-nuScenes validation set without using camera mask. The results show that our DAOcc made a relatively complete prediction of the scene, accurately reproducing it in fine detail.

5. Conclusion

In this paper, we propose a novel multi-modal occupancy prediction framework, aiming to achieve superior performance while using a deployment-friendly image feature extraction network and input image resolution. Extensive experiments on Occ3D-nuScenes and SurroundOcc demonstrate the superiority of our approach over existing methods. We believe that DAOcc can be put into practice and inspire future research.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 4

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 1, 5

[3] 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*, pages 11621–11631, 2020. 7

[4] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2, 5, 6

[5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 172–181, 2023. 4

[6] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023. 2

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1

[8] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A comprehensive framework for 3d occupancy estimation in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024. 6

[9] Adam W Harley, Shrinidhi K Lakshmikanth, Fangyu Li, Xian Zhou, Hsiao-Yu Fish Tung, and Katerina Fragkiadaki. Learning from unlabelled videos using contrastive predictive neural 3d mapping. *arXiv preprint arXiv:1906.03764*, 2019. 2, 4

[10] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 2, 4

[11] 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. 1, 4, 5, 7

[12] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 4, 6

[13] Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. *arXiv preprint arXiv:2311.07152*, 2023. 2, 3, 4

[14] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[15] 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

[16] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2561–2569, 2024. 2

[17] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 5

[18] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. 4

[19] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 17182–17191, 2022. 4

[20] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 2, 3, 4

[21] 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*, pages 1477–1485, 2023. 1, 5

[22] 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. 3, 6

[23] 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. 1, 5, 6

[24] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 3

[25] 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. 4, 5

[26] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023. 2

[27] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. *arXiv preprint arXiv:2402.03634*, 10, 2024.

[28] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 2

[29] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 3, 6, 7

[30] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2, 4

[31] Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. Lion: Linear group rnn for 3d object detection in point clouds. *arXiv preprint arXiv:2407.18232*, 2024. 2

[32] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[33] 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. 1, 3, 5, 6, 7

[34] 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. 1, 3, 5, 6

[35] 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. 6

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[37] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 6

[38] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 1, 5

[39] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. *arXiv preprint arXiv:2407.04049*, 2024. 1, 3, 5, 6

[40] Yining Shi, Kun Jiang, Ke Wang, Kangan Qian, Yunlong Wang, Jiusi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, and Diange Yang. Effocc: A minimal baseline for efficient fusion-based 3d occupancy network. *arXiv preprint arXiv:2406.07042*, 2024. 1, 3, 4, 5, 6, 7

[41] 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, 4, 5, 6, 7

[42] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 2442–2447. IEEE, 2017. 1

[43] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11794–11803, 2021. 4

[44] Lizi Wang, Hongkai Ye, Qianhao Wang, Yuman Gao, Chao Xu, and Fei Gao. Learning-based 3d occupancy prediction for autonomous navigation in occluded environments. In *2021 IEEE/RSJ International Conference on Intelligent*

Robots and Systems (IROS), pages 4509–4516. IEEE, 2021. 1

[45] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 1

[46] 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, 2, 4, 5, 6

[47] 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. 1, 3, 4, 5, 6

[48] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 2, 4, 6, 7

[49] 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. 1, 3, 6

[50] Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*, 2024. 1, 5, 6

[51] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 1

[52] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2, 4

[53] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 5

[54] 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. 1, 2, 3, 4, 6

[55] Zichen Yu, Changyong Shu, Qianpu Sun, Junjie Linghu, Xiaobao Wei, Jiangyong Yu, Zongdai Liu, Dawei Yang, Hui Li, and Yan Chen. Panoptic-flashocc: An efficient baseline to marry semantic occupancy with panoptic via instance center. *arXiv preprint arXiv:2406.10527*, 2024. 5, 6, 7

[56] Jinglin Zhan, Tiejun Liu, Rengang Li, Jingwei Zhang, Zhaoxiang Zhang, and Yuntao Chen. Real-aug: Realistic scene synthesis for lidar augmentation in 3d object detection. *arXiv preprint arXiv:2305.12853*, 2023. 2

[57] 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. 1, 6

[58] Ji Zhang and Yiran Ding. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction. *arXiv preprint arXiv:2403.05329*, 2024. 1, 3, 4, 5, 6

[59] 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. 2, 6

[60] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5

[61] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 7, 8

[62] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 1

[63] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3