OctOcc: High-Resolution 3D Occupancy Prediction with Octree

Wenzhe Ouyang[1]*, Xiaolin Song[2]†, Bailan Feng[2], Zenglin Xu[1,3]†

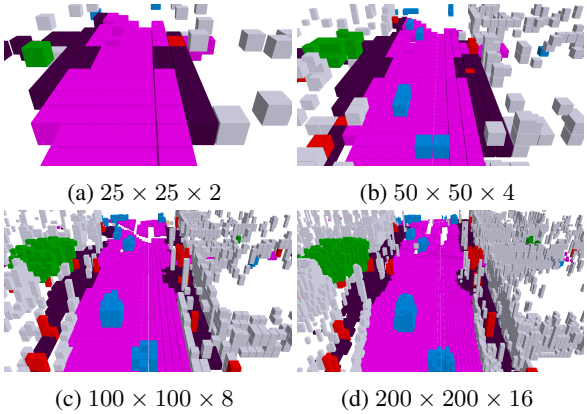
[1]Harbin Institute of Technology, Shenzhen, Guandong, China
[2]Huawei Noah's Ark Lab, Beijing, China
[3]Peng Cheng Lab, Shenzhen, Guandong, China
20B951020@stu.hit.edu.cn, songxiaolin2@huawei.com, fengbailan@huawei.com, zenglin@gmail.com

Abstract

3D semantic occupancy has garnered considerable attention due to its abundant structural information encompassing the entire scene in autonomous driving. However, existing 3D occupancy prediction methods contend with the constraint of low-resolution 3D voxel features arising from the limitation of computational memory. To address this limitation and achieve a more fine-grained representation of 3D scenes, we propose OctOcc, a novel octree-based approach for 3D semantic occupancy prediction. OctOcc is conceptually rooted in the observation that the vast majority of 3D space is left unoccupied. Capitalizing on this insight, we endeavor to cultivate memory-efficient high-resolution 3D occupancy predictions by mitigating superfluous cross-attentions. Specifically, we devise a hierarchical octree structure that selectively generates finer-grained cross-attentions solely in potentially occupied regions. Extending our inquiry beyond 3D space, we identify analogous redundancies within another side of cross attentions, 2D images. Consequently, a 2D image feature filtering network is conceived to expunge extraneous regions. Experimental results demonstrate that the proposed OctOcc significantly outperforms existing methods on nuScenes and SemanticKITTI datasets with limited memory consumption.

Introduction

3D scene understanding has been a pivotal and fundamental task in computer vision and autonomous driving systems for years. Earlier methodologies predominantly hinged upon LiDAR sensors to grapple with this challenge, which is limited to exorbitant hardware costs and the sparse scanned points clouds. Due to its inherent potential for developing cost-effective autonomous driving systems, the vision-centric perception has recently gained remarkable traction within both industry and academia. Taking multiple surrounding camera image as input, vision-centric models have evinced promising performance on various 3D scene understanding tasks such as 3D object detection (Wang et al. 2021b,c; Li et al. 2022b,a; Zhou et al. 2023; Li et al. 2023a), 3D map segmentation (Hu et al. 2021; Akan et al. 2022; Zhang et al. 2022), and depth estimation (Guizilini et al. 2022; Wei et al. 2022).



(a) $25 \times 25 \times 2$ (b) $50 \times 50 \times 4$ (c) $100 \times 100 \times 8$ (d) $200 \times 200 \times 16$

Figure 1. An illustration of different resolution of a scene representation: (a) $25 \times 25 \times 2$, (b) $50 \times 50 \times 4$, (c) $100 \times 100 \times 8$, and (d) $200 \times 200 \times 16$. The lower spatial resolution is hard to represent complex geometric shapes.

Despite the commendable achievements of vision-based models in the 3D object detection task, a conspicuous obstacle remains in their ability to faithfully encapsulate intricate geometric shapes and complex autonomous driving scenarios in the real world. This challenge arises from the conventionally employed coarse-level 3D bounding box representation of foreground objects, alongside the neglect of background elements. Recently, 3D occupancy prediction has been attended to as a promising fine-grained representational framework. This framework entails the allocation of semantic occupancy to each voxel in 3D space, thus engendering the pursuit of a granular depiction of 3D scenes. Several pioneering studies facilitate 3D occupancy prediction through the generation of 3D occupancy labels via LiDAR points (Tian et al. 2023; Sima et al. 2023; Wang et al. 2023), coupled with the formulation of baseline models (Huang et al. 2023; Wei et al. 2023; Zhang et al. 2023).

However, existing approaches are inclined to acquire low-resolution 3D voxel features, no matter they are network-based (Wei et al. 2023; Huang et al. 2023; Wang et al. 2023) or the geometry-based (Zhang et al. 2023). While these

*This work was done when Wenzhe Ouyang was an intern in Huawei Noah's Ark Lab.
†Corresponding author

low-resolution features are employed for producing the final high-resolution 3D occupancy predictions, trilinear interpolation or deconvolution operations are widely adopted in these works, which introduce challenges in terms of accurately predicting object shapes due to the attenuation of high-frequency details. More critically, it is challenging to represent the abundant semantic information within a low spatial resolution, as shown in Fig. 1, in which many important semantic details are ignored in low-resolution representation. Consequently, a substantial portion of prediction errors is traceable to either the insufficient spatial resolution or 3D occupancy supervision itself being performed at a coarse level. It naturally motivates us to develop a method to predict high-resolution 3D occupancy directly, which, however, leads to a rapid increase in computational consumption.

To address this issue, we propose OctOcc. OctOcc is conceptually rooted in the observation that there is no semantic information in the vast majority of the 3D space of autonomous driving scenarios, which results in a plethora of redundancy of 3D-2D spatial cross attention. Thus, our key insight is to reduce such redundancy to achieve a memory-efficient high-resolution 3D occupancy prediction:

- We first consider diminishing the redundancy in 3D voxels by a hierarchical octree structure. At the coarse level of octree structure, we focus on identifying occupied regions, regardless of the semantic information. In the neighbor finer layer, the OctOcc selectively generates octant sub-tree voxels to those nodes that are detected as potentially occupied regions only, and then applies finer-grained cross-attentions to those voxels. Our method outputs semantic categories to each voxel in the last layer to achieve high-resolution 3D occupancy predictions. Such a paradigm effectively suppresses the hunger for high-resolution 3D voxel queries for colossal memory consumption, thereby laying the foundation of the OctOcc to achieve high-resolution 3D occupancy prediction.
- We then further investigate the other side of 3D-2D spatial cross-attention, 2D image features. We found that there also exists much redundant or irrelevant information in 2D images, such as sky or far-away foreground objects. Therefore, we devise a filter mask prediction network to expunge those obviously irrelevant parts in 2D surrounding images.

To validate the effectiveness of the proposed method, we conduct comprehensive experiments on vision-based benchmarks, the surrounding view dataset nuScenes (Caesar et al. 2020) and monocular view SemanticKITTI (Behley et al. 2019). Experimental results show that the proposed OctOcc significantly improves the performance of 3D occupancy predictions(+3.0% mIoU) with a limited increase in memory consumption(less than 7GB) and computational costs. The qualitative results indicate that our method generates more fine-granularity 3D occupancy predictions.

Related Work

Vision-Centric Perception. Vision-centric 3D perception conducted in bird's eye-view (BEV) recently emerged as a promising alternative to the LiDAR-based autonomous driving solutions. It has achieved promising performance on several autonomous-driving related tasks, such as 3D object detection (Wang et al. 2021b,c; Li et al. 2022b,a; Zhou et al. 2023; Li et al. 2023a), map segmentation (Pan et al. 2020; Roddick et al. 2020; Saha et al. 2021; Zou et al. 2023; Saha et al. 2022; Gong et al. 2022; Zhou et al. 2022), and lane segmentation (Garnett et al. 2019; Guo et al. 2020; Chen et al. 2022; Liu et al. 2022a). These works can be divided into two main categories based on 3D-2D view transformation: geometry-based and network-based. The geometry-based methods (Philion et al. 2020; Huang et al. 2021, 2022; Li et al. 2023a) fully utilize the geometric relationship of the camera to lift 2D features to 3D space by explicit or implicit depth estimation. The network-based methods (Wang et al. 2021b; Liu et al. 2022b; Li et al. 2022b) employ a top-down strategy by directly constructing BEV queries and searching corresponding features on front-view images by the cross-attention mechanism. Though existing works have achieved competitive performance on 3D object detection tasks, they still suffer from the coarse-level 3D bounding box representation, which limits its application on fine-grained 3D scene representation.

Voxel-Based Scene Reconstruction. 3D scene reconstruction is a fundamental but challenging task in computer vision. Voxel-based scene reconstruction voxelized the 3D space into discretized voxels and described each voxel by a feature vector. The ability to describe fine-granularity 3D scenes makes voxel-based scene reconstruction favorable for 3D scene understanding tasks such as lidar segmentation (Cheng et al. 2021; Ye et al. 2023) and 3D scene completion (Yan et al. 2021; Cao et al. 2022). Though these methods achieved success in 3D scene understanding, they are usually in a LiDAR-centric paradigm. Besides, 3D scene reconstruction methods (Murez et al. 2020; Sun et al. 2021; Bozic et al. 2021) reconstruct accurate 3D geometry and scene in a vision-centric paradigm. However, most of these methods are designed for indoor scenes, which is quite different from outdoor settings in autonomous driving scenarios. To the best of our knowledge, MonoScene (Cao et al. 2022) is the first work to reconstruct outdoor scenes, but it is tailored for monocular image input.

3D Occupancy Prediction. 3D occupancy prediction aims to reconstruct voxelized 3D scenes, which is similar to Occupancy Grid Mapping(OGM), a classical task in robotics. However, 3D occupancy prediction usually utilizes RGB images from surrounding cameras as input, but OGM often requires measurement from range sensors like LiDAR and RADARs. Tesla is the first to project the perspective view features onto the 3D voxels space to achieve the 3D occupancy prediction network. Subsequent endeavors, exemplified by OpenOccupancy (Wang et al. 2023), SSCBench (Li et al. 2023b), and Occ3D (Tian et al. 2023), concentrate on the construction of datasets or benchmarks for 3D occupancy predictions. TPVFormer (Huang et al. 2023) proposes a tri-perspective view method to predict 3D occupancy. SurroundOcc (Wei et al. 2023) designs a coarse-to-fine architecture with generated dense 3D occupancy as supervision. OccFormer (Zhang et al. 2023) employs a geometry-based paradigm to construct 3D voxel features. Despite the differ-

ence in 3D-2D view transformation, all existing works learn the 3D voxel features to depict the corresponding scenes. Therefore, the spatial resolution of the 3D voxel features is a non-negligible parameter for 3D Occupancy Prediction. However, existing works suffer from the low resolution of 3D voxel features due to limited memory and computation resources.

Method

OctOcc is a memory-efficient 3D occupancy prediction network. The key insight in our method is to diminish the redundancy in 3D-2D spatial cross-attention. On the 3D side, we construct a hierarchical octree structure for 3D voxels, in which we selectively deploy cross-attentions to those octant sub-tree nodes that detect as potentially occupied regions. On the 2D side, we design a 2D image feature filter network to filtrate those irrelevant images formation.

Preliminaries

3D Occupancy Prediction. Given a sequence of sensor inputs, the goal of 3D occupancy prediction is to estimate the state of each voxel, including occupancy("occupied", "free") and semantics (category or "other"). Formally, the 3D occupancy prediction can be represented as:

$$\mathbf{Occ} = \mathcal{F}(I^1, I^2, ...I^N), \tag{1}$$

where $\mathcal{F}$ is an neural network and $I^N \in \mathbb{R}^{H\times W\times 3}$ is input surrounding RGB images. $\mathbf{Occ} \in \mathbb{R}^{H\times W\times L}$ is the 3D occupancy predictions. Typically, the 3D occupancy task assumes that some sensor intrinsic parameters K_i and extrinsic parameter $[R_i|t_i]$ are known.

3D occupancy is a good representation of multi-camera 3D scene reconstruction. First, 3D occupancy provides more intricate geometric structures of objects compared to 3D object detection. Second, 3D occupancy can easily extend to further downstream tasks, such as 3D semantic segmentation and scene flow estimation.

3D-2D View Transformation. To map the homography between 2D perspective view with 3D feature space, transformers with cross attention is a prevalent choices for the network-based methods. 3D voxel queries aggregate 2D image features into 3D space via following 3D-2D spatial cross-attention operation (Li et al. 2022b):

$$\mathbf{SCA}(\mathbf{Q}_p^l, \mathbf{F}) = \sum_{j=1}^{N_{ref}} \mathbf{DeformAttn}(\mathbf{Q}_p^l, P_{(p,i,j)}, \mathbf{F}^i), \tag{2}$$

where i and j indexes the camera view and the reference points, l indexes the layer. N_{ref} is the total reference points for each queries, and $P_{(p,i,j)}$ is implemented to obtain the j-th reference point on the i-th view image. F^i is the 2D image features of the i-th camera view.

Octree for 3D Occupancy Predictions

In this section, we introduce our octree-based approach for memory-efficient high-resolution 3D occupancy predictions. To formalize our approach, we first introduce the definition of "occupied region" and its statistical analysis. Then, we describe the proposed octree-based method to detect occupied regions and predict fine-grained 3D occupancy.

Resolution	200	100	50	25
Occupied	4.88%	10.81%	12.44%	13.68%
Unoccupied	95.12%	89.19%	85.56%	86.32%

Table 1. Statistical analysis of the occupied regions in different spatial resolutions. 200, 100, 50 and 25 represent the resolution with $200 \times 200 \times 16$, $100 \times 100 \times 8$, $50 \times 50 \times 4$ and $25 \times 25 \times 2$, respectively.

Occupied Regions. Intuitively, the majority of the semantic information in an autonomous driving scene concentrates on the lower half of the surrounding 3D space, close to the drivable surfaces. Based on this intuition, we conduct a statistical analysis on the occupied areas at different resolutions, as shown in the Table 1. Here, we define the voxel space where semantic information occupies more than 10% of the space as **occupied area**.

As shown in Table 1, it is apparent that a substantial portion of space in the autonomous driving scenes actually are unoccupied. Furthermore, with the increases in 3D spatial resolution, fewer voxels are occupied by semantic information, even less than 5% in $200 \times 200 \times 16$ resolution. This revelation led us to consider that a majority of regions in the 3D space do not require dense and high-resolution 3D cross-attentions, given the scarcity of semantic information in those areas. By confining the utilization of high-resolution cross-attentions to partial regions, we can significantly diminish memory consumption. Building upon this observation, we propose OctOcc, which implements fine-grained 3D-2D cross-attention exclusively within potentially occupied regions.

Octree Definition and Construction. An overview of our Octree-based 3D occupancy predictions network is shown in Fig. 2. First, 2D image features are extracted from multi-view images with an image backbone. The 3D voxel queries utilize a hierarchical tree-like design, and we define a node $octree$ to represent a 3D voxel. A gray voxel in Fig. 2 represents detecting potentially occupied regions. As shown in Fig. 2, each occupied voxel in the coarse levels has octant child voxels in its neighboring lower-level voxel queries. Potential occupied regions in each layer are used to construct a hierarchical octree. Each node query aggregates multi-level 2D image features into 3D space via a 3D-2D spatial cross-attention operation.

After defining the octree, we start from the roughest layer, all root nodes in this layer will be used for cross-attentions queries by default. Next, we can generate updated 3D voxels with the coarsest level by feeding these queries into the 3D-2D cross-attentions. We then upsample the updated 3D voxel features using trilinear interpolation. Different from SurroundOcc (Wei et al. 2023) supervised with coarse-level semantic labels, we feed those upsampling 3D voxels features into a binary classifier to predict potential occupied regions. Here, we use binary occupancy labels as supervision, generating these labels according to the definition in Sec.

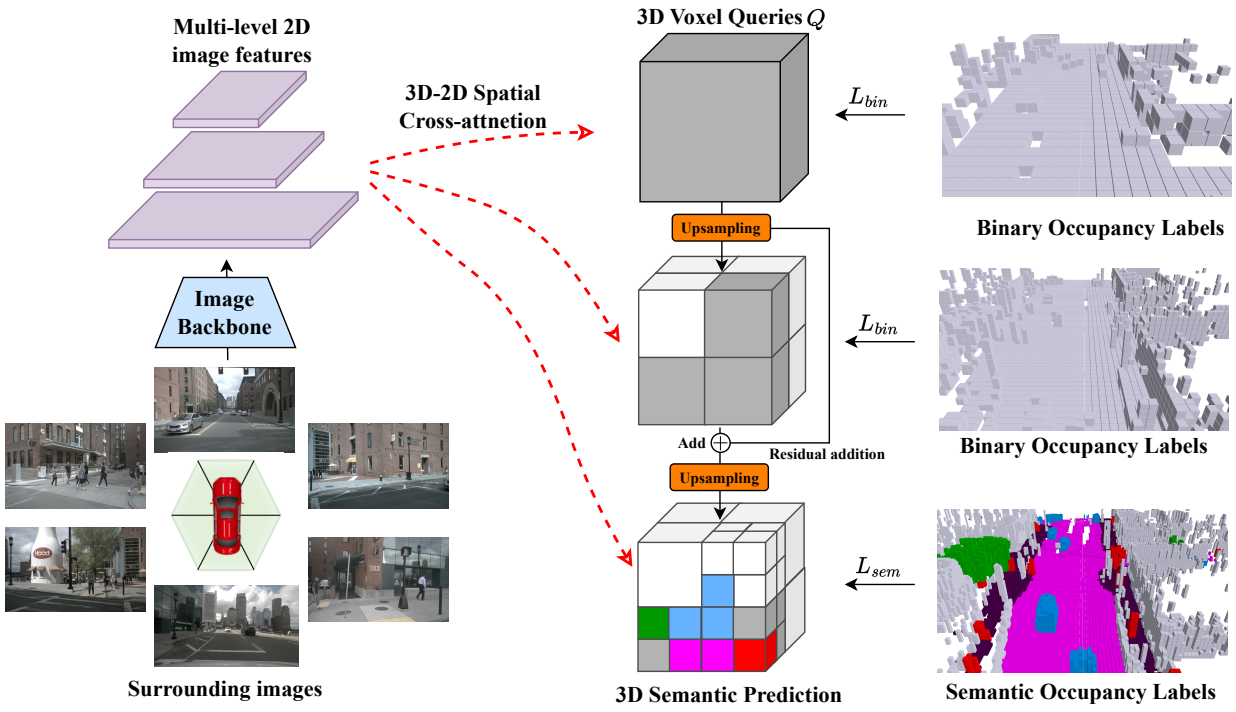


Figure 2. An overview of the proposed octree-based 3D occupancy prediction network. The octree-based structure identifies occupied regions at the coarse levels and generates octant sub-tree queries in these regions. By that, these limited sub-tree queries are subjected to fine-granularity cross attention at the finer level. Finally, our method predicts semantic categories for each voxel in the last layer to achieve high-resolution 3D occupancy predictions.

Occupied regions. The predicted top K voxels are considered to be potentially occupied regions, and are selected for indexing child nodes on the finer layer with higher spatial resolution:

$$\mathbf{Q}^{l+1}_{occupied} = \mathbf{SCA}(\mathbf{top}_k(\mathbf{Upsample}(\mathbf{Q}^l), \mathbf{F})), \quad (3)$$

where $\mathbf{Q}^l$ and $\mathbf{Q}^{l+1}$ represent two neighbor 3D voxel queries, and *occupied* represent the indexes of selected top K voxels. Once we obtain updated occupied 3D voxel features, we proceed to update the whole 3D voxel features with the following:

$$\mathbf{Q}^{l+1}_{unoccupied} = \mathbf{Upsample}(\mathbf{Q}^l), \quad (4)$$

$$\mathbf{Q}^{l+1} = \mathbf{Concat}(\mathbf{Q}^{l+1}_{occupied}, \mathbf{Q}^{l+1}_{unoccupied}), \quad (5)$$

where $\mathbf{Upsample}()$ represent a trilinear interpolation operation and $\mathbf{Concat}$ represent a concatenate operation. After obtaining the whole updated 3D voxel features, we apply a 3D convolutional layer to enhance feature interaction throughout the entire 3D feature voxels. Finally, we apply two linear layers with softplus activation (Zheng et al. 2015) to predict binary or multi-class semantic predictions:

$$\mathbf{Occ}^{l+1} = \mathbf{Linear}(\mathbf{Softplus}(\mathbf{Linear}(\mathbf{Q}^{l+1}))). \quad (6)$$

Here, we regress 3D multi-class semantic predictions in the last layer only, as shown in Fig. 2.

2D Image Feature Filtering

The hierarchical octree is designed to diminish redundant 3D voxel queries in 3D-2D cross attention (Li et al. 2022b). Additionally, we further investigate another side of the 3D-2D spatial cross attention, 2D image features. BEVFormer (Li et al. 2022b) and the following works idiomatically consider all 3D voxels which are initially projected in the 2D image with perspective transformation. However, we find the 2D RGB image also contains much redundant information, as shown in the upper left corner of Fig. 3, which includes much irrelevant background information (sky, etc.) and far-away foreground objects (over 51.2 meters). To address this, we propose a 2D image feature filter to eliminate redundant 3D-2D cross-attention which initially projected on these irrelevant regions:

$$P_{(p,i,j)} \in \mathbf{Filter}(\mathbf{F}^i), \quad (7)$$

where $P_{(p,i,j)}$ is the reference points in Equation (2).

As shown in Fig. 3, we found that these irrelevant regions in 2D images are strongly related to the projection depth of LiDAR due to the installation location and operating mechanism of the LiDAR. Therefore, we directly apply a binarization and dilation operation to generate the corresponding dilated depth masks as supervision. We utilize a 2D convolution layer to predict a binary filter mask.

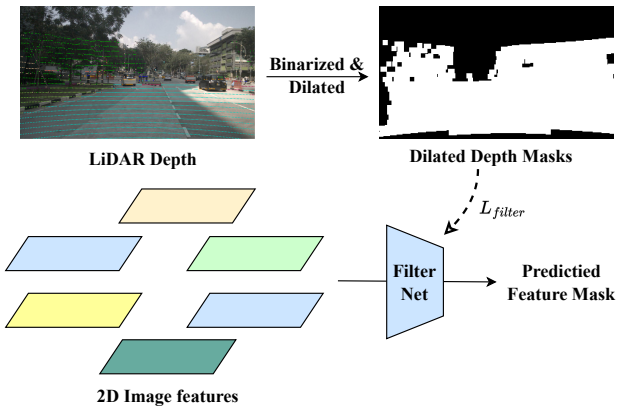


Figure 3. An illustration of 2D Feature Mask Filtering. To generate 2D mask labels, we binarize and dilate the projected LiDAR depth information. By learning from these dilated masks, the filter network effectively filters out irrelevant 2D regions in the images.

Training Loss

Existing works widely utilize cross-entropy loss as supervision for 3D occupancy predictions. To reduce the influence of the class-imbalance, the class-weighted cross-entropy loss is further utilized in (Zhang et al. 2023; Li et al. 2023c). Although effective, the class-weighted cross-entropy loss has its inherent shortcomings, and performance improvement is also limited. Thus, we consider Generalized Dice loss (Sudre et al. 2017) as extra supervision in this paper:

$$\mathbf{L}_{dice} = 1 - 2\frac{\sum_{l=1}^{n} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{n} w_l \sum_n (r_{ln} + p_{ln})}, \tag{8}$$

where r_{ln} represents the ground truth of class l in position n, and p_{ln} represents the prediction softmax probability of class l in position n. w_l represent the weight of class l. We experimentally found that the original implementation of w_l is not ideal since the original implementation in medical scenes (Sudre et al. 2017) tends to underweight those semantic classes with high frequency in autonomous driving. Therefore, we modified the w_l as:

$$w_l = \frac{1}{\sum_{i=1}^{n} r_{ln}}, \tag{9}$$

where the weight w_l is inversely proportional to the sum of ground truth pixels rather than squared. The semantic occupancy training loss consists two parts:

$$\mathbf{L}_{occ} = \lambda_{ce}\mathbf{L}_{ce} + \mathbf{L}_{dice}, \tag{10}$$

where λ_{ce} is balancing coefficients, and we set $\lambda_{ce} = 2$ in this paper. We use binary cross-entropy loss for coarse binary occupancy prediction and feature mask filtering loss. The final loss is formulated as follows:

$$L = \lambda_{occ}\mathbf{L}_{occ} + \lambda_{filter}\mathbf{L}_{filter} + \mathbf{L}_{bin}, \tag{11}$$

where λ_{occ} and λ_{filter} are balancing coefficients, both set to 2 in this paper.

Experiments

Experimental Setting

Datasets and Evaluation Metric. Following existing works (Huang et al. 2023; Wei et al. 2023; Zhang et al. 2023), we conduct experiments on the nuScenes dataset (Caesar et al. 2020), a large-scale autonomous driving dataset. We use the surrounding 6 RGB images of nuScenes as the input, and the input image resolution is resized to 1280 $\times$ 720. We utilize the 3D occupancy labels generated from Occ3D(Tian et al. 2023). The occupancy prediction range is set as $[-40m, 40m]$ for the X and Y axis and $[-1m, 5.4m]$ for the Z axis. The final output occupancy has the resolution with $200 \times 200 \times 16$, and the voxel size is 0.4m.

To further demonstrate the effectiveness of our method, we conduct a monocular semantic scene completion experiment on SemanticKITTI (Behley et al. 2019). SemanticKITTI has annotated outdoor LiDAR scans with 21 semantic labels. The ground truth is voxelized as $256 \times 256 \times 32$ grid with 0.2m voxel size. We evaluate our model on the validation set.

For 3D semantic occupancy prediction, we use mean Intersection over Union (mIoU) to evaluate the performance of a model. For the monocular scene semantic completion task, we follow (Li et al. 2023c) to use mIoU and Intersection over Union (IoU) as metrics.

Implementation Details. For the 3D occupancy prediction task in nuScenes dataset (Caesar et al. 2020), we utilize feature maps $C_3 \sim C_5$ (i.e., three scales) from the backbone network following SurroundOcc (Wei et al. 2023). For a fair comparison, we use ResNet101-DCN (He et al. 2015) with initial weight from FCOS3D (Wang et al. 2021a) as the backbone to extract image features. The whole network architecture is set to 4 levels, and the resolution of 3D voxel queries is set to $25 \times 25 \times 2$, $50 \times 50 \times 4$, $100 \times 100 \times 8$, $200 \times 200 \times 16$. The first three layers are supervised with binary 3D occupancy labels, and the last layer is supervised with 3D semantic occupancy labels. The values of top k between four 3D voxel features layers are set to 625, 3000, and 15000, respectively.

For the semantic scene completion task in SemanticKITTi dataset (Behley et al. 2019), we follow MonoScene (Cao et al. 2022) and use the ResNet-50 (He et al. 2015) as the backbone. We adopt FPN (Lin et al. 2017) to further fuse the feature of different levels for both tasks.

Main Results

3D Occupancy Predictions. We first compare the proposed OctOcc with other methods on the 3D semantic occupancy prediction task. Due to the differences in the 3D occupancy labels of existing methods, we re-trained BEVFormer (Li et al. 2022b), TPVFormer (Huang et al. 2023), and SurroundOcc (Wei et al. 2023) under the same 3D occupancy labels (Tian et al. 2023) for a fair comparison. As shown in Table 2, the proposed OctOcc outperforms all other competing methods with a large margin(almost **+3.0**). We also show some qualitative results in Fig. 4. The first and second rows of Fig. 4 demonstrate the superiority of the proposed method

Method	SSC mIoU↑	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	man-made	vegetation
BEVFormer (Li et al. 2022b)	24.97	6.58	36.4	14.12	37.46	42.33	13.33	20.29	17.76	16.35	16.2	29.99	52.95	30.53	31.68	25.85	16.94	15.77
TPVFormer (Huang et al. 2023)	23.25	7.25	37.19	19.47	39.19	43.61	15.39	21.42	22.31	22.98	20.35	30.68	55.92	28.04	33.03	27.79	15.83	17.0
SurroundOcc (Wei et al. 2023)	27.03	6.24	36.98	17.13	**41.0**	43.97	15.67	19.85	15.16	11.51	24.57	31.12	56.94	31.48	36.92	32.61	19.51	18.89
Occ3D (Tian et al. 2023)	28.53	8.09	39.33	20.56	38.29	42.24	**16.93**	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.0
OctOcc	**31.52**	**11.1**	**42.83**	**20.9**	37.63	**44.26**	12.6	**25.57**	**23.07**	**25.38**	**26.95**	**31.25**	**66.64**	**36.05**	**40.57**	**38.98**	**25.24**	**23.87**

Table 2. 3D semantic occupancy prediction results on nuScenes validation set. Due to the differences in the 3D occupancy labels of existing methods, we re-trained other models under the same 3D occupancy labels for a fair comparison.

Method	mIoU↑	IoU↑	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
MonoScene (Cao et al. 2022)	11.08	36.86	56.5	26.7	14.3	0.5	14.1	23.3	7.0	0.6	0.5	1.5	17.9	2.8	29.6	1.9	1.2	0.0	5.8	4.1	2.3
TPVFormer (Huang et al. 2023)	11.36	35.61	56.5	25.9	20.6	**0.9**	13.9	23.8	8.1	0.4	0.0	4.4	16.9	2.3	30.4	0.5	0.9	0.0	5.9	3.1	1.5
VoxFormer (Li et al. 2023c)	13.35	44.15	53.6	26.5	19.7	0.4	19.5	26.5	7.3	1.3	0.6	**7.8**	26.1	6.1	33.0	1.9	2.0	0.0	7.3	9.2	4.9
OccFormer (Zhang et al. 2023)	13.46	36.50	**58.9**	26.9	19.6	0.3	14.4	25.1	**25.5**	0.8	1.2	8.5	19.6	3.9	32.6	2.8	**2.8**	0.0	5.6	4.3	2.9
OctOcc	**14.59**	**44.02**	55.1	**27.9**	**22.6**	0.5	**20.3**	**27.8**	6.0	**2.6**	**2.0**	6.8	**26.6**	**6.8**	**33.8**	**2.7**	0.0	0.0	**8.9**	**9.3**	**5.6**

Table 3. 3D Semantic scene completion results on SemanticKITTI validation set. All experiments are conducted under the resolution with $256 \times 256 \times 32$, and the results are reported in VoxFormer (Li et al. 2023c) and OccFormer (Zhang et al. 2023).

Method	Resolution	Memory	mIoU↑
BEVFormer (Li et al. 2022b)	200×200	14.4G	24.97
TPVFormer (Huang et al. 2023)	$100 \times 100 \times 8$	23.9G	23.25
SurroundOcc (Wei et al. 2023)	$100 \times 100 \times 8$	16.4G	27.03
OccFormer (Zhang et al. 2023)	$100 \times 100 \times 8$	Over 32G	-
w/o Octree	$100 \times 100 \times 8$	16.6G	29.51
w/o Octree	$200 \times 200 \times 16$	Over 32G	-
OctOcc	$200 \times 200 \times 16$	23.4G	**31.52**

Table 4. Ablation studies on the Octree structure design. The proposed octree design significantly reduces the memory consumption of high-resolution 3D occupancy predictions and brings a considerable performance improvement at the same time.

in the more accurate detection of foreground objects, such as cars and traffic cones. Besides, Fig. 4 also shows that our method has a fine-grained geometric shape on background information, an obvious advantage of high-resolution representation.

Semantic Scene Completion. To further demonstrate the effectiveness of our method, we also conduct monocular 3D semantic scene completion on SemanticKITTI (Behley et al. 2019). As shown in Table 3, our method also achieves state-of-the-art performance on this benchmark, even though our method is not designed for monocular perception.

More visualization results and videos can be found in the supplementary material.

Method	mIoU(8 epoch)↑	mIoU(24 epoch)↑
w/o 2D Feature Filtering	29.8	31.23
OctOcc	**31.1(+1.3)**	**31.52(+0.3)**

Table 5. Ablation studies on the 2D Image feature filtering. The proposed 2D Feature Filtering converges obviously faster while improving the performance of 3D occupancy prediction.

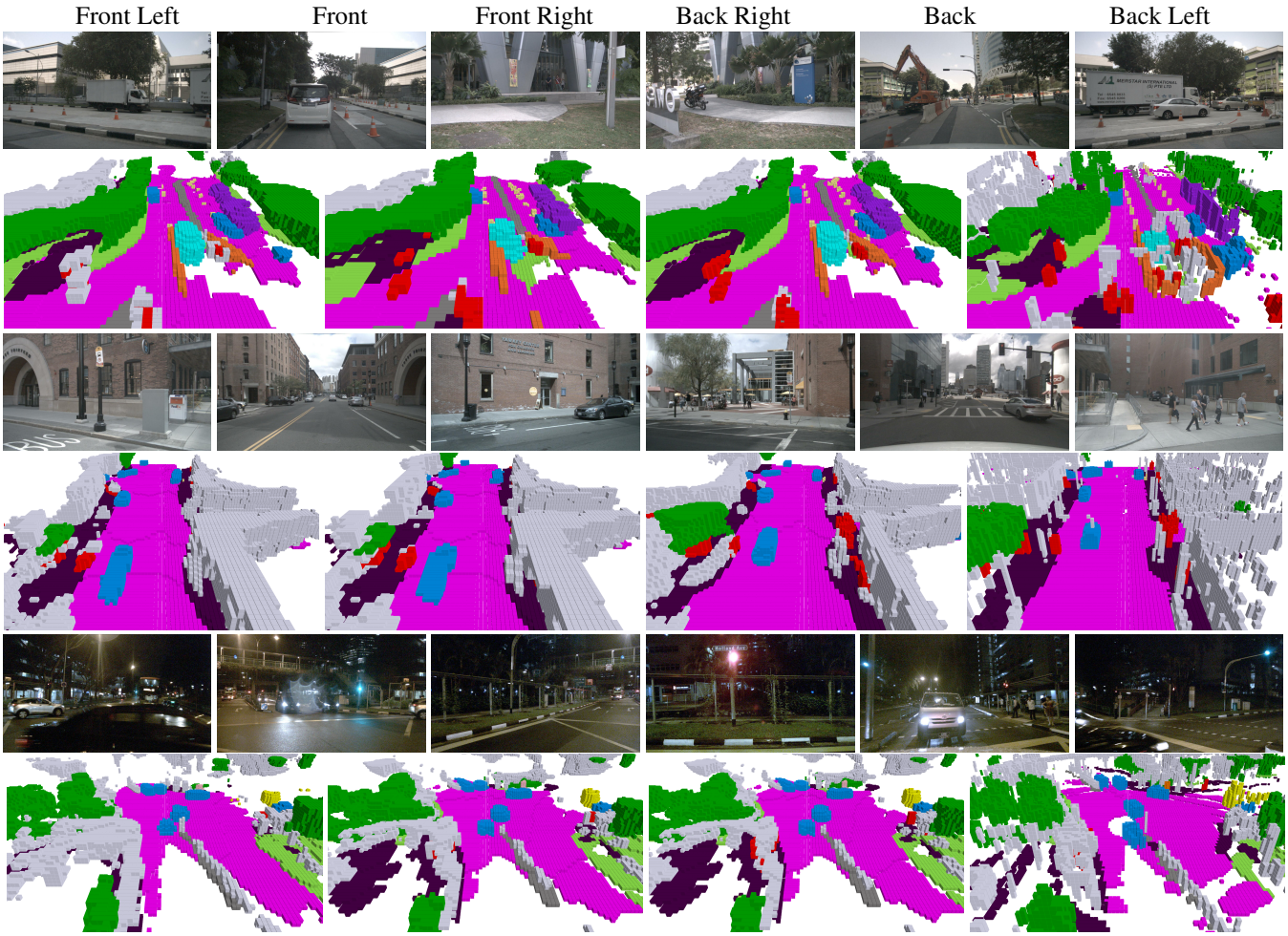
Method	mIoU↑
CE loss only	29.32
CE + Sem&Geo loss (Cao et al. 2022)	31.06
CE + Gene. Dice loss (Sudre et al. 2017)	30.64
CE + modified Gene. Dice loss	**31.52**

Table 6. Ablation studies on the loss design. The proposed modified Generalized Dice loss performs best, and the original design of Generalized Dice loss (Sudre et al. 2017) even degrades the performance.

Ablation Studies

The ablation is conducted on nuScenes dataset (Caesar et al. 2020) and from three perspectives: the octree structure design, the 2D image feature filtering, and the loss design.

Octree Structure Design. Table 4 ablates the octree structure design for 3D occupancy predictions. As shown in Table 4, the octree design can effectively reduce memory consumption to generate high-resolution 3D voxel features. As



(a) BEVFormer (Li et al. 2022b) (b) SurroundOcc (Wei et al. 2023) (c) OctOcc(Ours) (d) Ground Truth

Figure 4. Qualitative results of our method and others. The odd rows represent the six surrounding input images, and the even rows represent the corresponding 3D occupancy predictions. The OctOcc better captures foreground objects, such as cars, pedestrians, and traffic cones. Furthermore, OctOcc shows finer-granularity geometric shapes on the whole scene.

a comparison, previous methods could not generate high-resolution 3D voxel features with less than 32G memory, no matter the network-based (Huang et al. 2023; Wei et al. 2023) or the geometry-based (Zhang et al. 2023). The proposed octree design produces **+2.01** improvement with 6.8G extra memory consumption.

2D Image Features Filtering. In Table 5, we ablate the proposed filtering mechanism for 2D image features, which aims to eliminate redundant 3D-2D spatial cross-attention further. Thanks to the 2D image feature filtering design, the proposed method can achieve faster convergence with 1/3 of training epochs. At the same time, this design can also slightly enhance the performance.

Loss Design. Table 6 compares different loss designs for the 3D occupancy predictions. As shown in Table 6, using naive cross-entropy loss obviously lag behind others due to the imbalance of semantic information. The proposed modified Generalized Dice loss achieves the best performance, even compared with scene-class affinity loss in MonoScene (Cao et al. 2022). Table 6 also shows the original design of Generalized Dice (Sudre et al. 2017) cannot work ideally in 3D occupancy predictions.

Conclusion

In this paper, we have presented OctOcc, a high-resolution 3D occupancy prediction method with Octree. With the insight for diminishing those redundant 3D-2D cross-attentions, we propose a hierarchical octree structure to deploy finer-granularity cross-attentions exclusively in those potentially occupied regions, which significantly reduces memory consumption. Additionally, we design a filtering network to reduce redundancy in 2D image features. OctOcc has achieved state-of-the-art performance for 3D occupancy predictions on nuScenes and semantic scene completion on SemanticKITTI. We hope that the proposed octree structure and the findings about redundancy in cross-attentions will be helpful to other 3D voxel prediction tasks.

Acknowledgements

This work was partially supported by a key program of fundamental research from Shenzhen Science and Technology Innovation Commission (No.JCYJ20200109113403826), the Major Key Project of PCL (No. PCL2021A06), and an Open Research Project of Zhejiang Lab (NO.2022RC0AB04).

References

Akan, A. K.; and Güney, F. 2022. StretchBEV: Stretching Future Instance Prediction Spatially and Temporally. *European Conference on Computer Vision (ECCV)*.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*.

Bozic, A.; Palafox, P.; Thies, J.; Dai, A.; and Niessner, M. 2021. TransformerFusion: Monocular RGB Scene Reconstruction using Transformers. *Proc. Neural Information Processing Systems (NeurIPS)*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, A.-Q.; and Raoul, d. C. 2022. MonoScene: Monocular 3D Semantic Scene Completion. *CVPR*.

Chen, L.; Sima, C.; Li, Y.; Zheng, Z.; Xu, J.; Geng, X.; Li, H.; He, C.; Shi, J.; Qiao, Y.; and Yan, J. 2022. PersFormer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark. *European Conference on Computer Vision (ECCV)*.

Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; and Liu, B. 2021. (AF)2-S3Net: Attentive Feature Fusion With Adaptive Feature Selection for Sparse Semantic Segmentation Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Garnett, N.; Cohen, R.; Pe'er, T.; Lahav, R.; and Levi, D. 2019. 3D-LaneNet: End-to-End 3D Multiple Lane Detection.

Gong, S.; Ye, X.; Tan, X.; Wang, J.; Ding, E.; Zhou, Y.; and Bai, X. 2022. GitNet: Geometric Prior-based Transformation for Birds-Eye-View Segmentation. *European Conference on Computer Vision (ECCV)*.

Guizilini, V.; Vasiljevic, I.; Ambrus, R.; Shakhnarovich, G.; and Gaidon, A. 2022. Full Surround Monodepth From Multiple Cameras. *IEEE Robotics and Automation Letters*, 7(2): 5397–5404.

Guo, Y.; Chen, G.; Zhao, P.; Zhang, W.; Miao, J.; Wang, J.; and Choe, T. E. 2020. Gen-lanenet: A Generalized and Scalable Approach for 3D lane detection. *European Conference on Computer Vision (ECCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

Hu, A.; Murez, Z.; Mohan, N.; Dudas, S.; Hawke, J.; Badrinarayanan, V.; Cipolla, R.; and Kendall, A. 2021. FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15273–15282.

Huang, J.; and Huang, G. 2022. BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; Yun, Y.; and Du, D. 2021. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View. *arXiv preprint arXiv:2112.11790*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2022a. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023a. BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection. *AAAI 2023*.

Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; Wang, Y.; Zhao, H.; Yu, Z.; and Feng, C. 2023b. SSCBench: A Large-Scale 3D Semantic Scene Completion Benchmark for Autonomous Driving. *arXiv preprint arXiv:2306.09001*.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023c. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. *European Conference on Computer Vision (ECCV)*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, R.; Chen, D.; Liu, T.; Xiong, Z.; and Yuan, Z. 2022a. Learning to predict 3d lane shape and camera pose from a single image via geometry constraints. *AAAI 2022*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022b. Petr: Position embedding transformation for multi-view 3d object detection. *European Conference on Computer Vision (ECCV)*.

Murez, Z.; van As, T.; Bartolozzi, J.; Sinha, A.; Badrinarayanan, V.; and Rabinovich, A. 2020. Atlas: End-to-End 3D Scene Reconstruction from Posed Images.

Pan, B.; Sun, J.; Leung, H. Y. T.; Andonian, A.; and Zhou, B. 2020. Cross-View Semantic Segmentation for Sensing Surroundings. *IEEE Robotics and Automation Letters*, 5(3): 4867–4873.

Philion, J.; and Fidler, S. 2020. Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D.

Roddick, T.; and Cipolla, R. 2020. Predicting Semantic Map Representations From Images Using Pyramid Occupancy Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saha, A.; Mendez, O.; Russell, C.; and Bowden, R. 2021. Enabling spatio-temporal aggregation in Birds-Eye-View Vehicle Estimation. *2021 IEEE International Conference on Robotics and Automation (ICRA)*.

Saha, A.; Mendez, O.; Russell, C.; and Bowden, R. 2022. Translating Images into Maps. *2022 International Conference on Robotics and Automation (ICRA)*.

Sima, C.; Tong, W.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; and Li, H. 2023. Scene as Occupancy. *International Conference on Computer Vision(ICCV), 2023*.

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Cardoso, M. J. 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. 240–248.

Sun, J.; Xie, Y.; Chen, L.; Zhou, X.; and Bao, H. 2021. NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video. *CVPR*.

Tian, X.; Jiang, T.; Yun, L.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. *arXiv preprint arXiv:2304.14365*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021a. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023. OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception. *International Conference on Computer Vision(ICCV), 2023*.

Wang, Y.; Guizilini, V.; Zhang, T.; Wang, Y.; Zhao, H.; ; and Solomon, J. M. 2021b. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. *The Conference on Robot Learning (CoRL)*.

Wang, Y.; and Solomon, J. M. 2021c. Object DGCNN: 3D Object Detection using Dynamic Graphs. *2021 Conference on Neural Information Processing Systems (NeurIPS)*.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Rao, Y.; Huang, G.; Lu, J.; and Zhou, J. 2022. SurroundDepth: Entangling Surrounding Views for Self-Supervised Multi-Camera Depth Estimation. *In CoRL*.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. *International Conference on Computer Vision(ICCV), 2023*.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ye, D.; Zhou, Z.; Chen, W.; Xie, Y.; Wang, Y.; Wang, P.; and Foroosh, H. 2023. LidarMultiNet: Towards a Unified Multi-Task Network for LiDAR Perception. *AAAI 2023*.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction.

Zhang, Y.; Zhu, Z.; Zheng, W.; Huang, J.; Huang, G.; Zhou, J.; and Lu, J. 2022. BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving. *arXiv preprint arXiv:2205.09743*.

Zheng, H.; Yang, Z.; Liu, W.; Liang, J.; and Li, Y. 2015. Improving deep neural networks using softplus units.

Zhou, B.; and Krähenbühl, P. 2022. Cross-View Transformers for Real-Time Map-View Semantic Segmentation. 13760–13769.

Zhou, H.; Ge, Z.; Li, Z.; and Zhang, X. 2023. MatrixVT: Efficient Multi-Camera to BEV Transformation for 3D Perception. *arXiv preprint arXiv:2211.10593*.

Zou, J.; Zhu, Z.; Huang, J.; Yang, T.; Huang, G.; and Wang, X. 2023. HFT: Lifting Perspective Representations via Hybrid Feature Transformation for BEV Perception. *2023 IEEE International Conference on Robotics and Automation (ICRA)*.