

# UnsOcc: 3D Semantic Occupancy Prediction in Unstructured Scene via Rendering Fusion

Ye Wu<sup>1\*</sup>, Ruiqi Song<sup>2,3\*</sup>, Baiyong Ding<sup>2,3</sup>, Nanxin Zeng<sup>1</sup>, Junjie Cheng<sup>1</sup> and Yunfeng Ai<sup>1,3†</sup>

**Abstract**—Unstructured scenes present unique challenges for autonomous driving, as irregular obstacles and sparse scene layouts undermine the effectiveness of traditional perception methods such as 3D object detection. 3D semantic occupancy prediction has emerged as a prominent focus due to its ability to provide dense spatial representations by assigning semantic labels to individual voxels in 3D space. However, directly applying 3D semantic occupancy prediction to unstructured scenes remains challenging because scene sparsity hinders effective cross-modal fusion and the more severe long-tail distribution in these scenarios further degrades prediction performance. To validate the effectiveness of our approach, we construct a dedicated dataset of unstructured scenes collected from open-pit mines. Based on this, we propose UnsOcc, a multi-modal 3D semantic occupancy prediction framework that improves robustness in unstructured environments. At its core, we introduce a rendering-based fusion module, RenderFusion, which enhances cross-modal feature alignment through bidirectional rendering supervision. Furthermore, we propose GSRefinement, a detail-aware auxiliary supervision method based on Gaussian Splatting that projects sparse 3D occupancy predictions into dense 2D semantic segmentation maps, enabling effective supervision for long-tail categories. Extensive experiments on both the open-pit mine dataset and the nuScenes dataset demonstrate that our method significantly outperforms existing state-of-the-art approaches.

## I. INTRODUCTION

With the continuous advancement of autonomous driving technology, its application scenarios have progressively extended to more complex unstructured scenes, such as the surface mine. However, traditional perception methods, such as 3D object detection [1], [2], often suffer performance degradation when confronted with irregular and long-tail obstacles, limiting their applicability in practical applications. To address these challenges, 3D occupancy prediction has emerged as a promising alternative. Rather than relying on 3D bounding boxes, this method divides 3D space into voxel grids and assigns a semantic label to each voxel, giving an efficient 3D semantic representation for unstructured scenes.

This work was supported by the Key Research and Development Program of Shaanxi Province under Grant 2024CY2-GJHX-49 and the Industry-University-Research Innovation Fund for Chinese Universities under Grant 2024HT023.

<sup>1</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China {wuye23, zengnanxin24, chengjunjie25}@mailsucas.ac.cn, aiyunfeng@ucas.ac.cn

<sup>2</sup>Institute of Automation, Chinese Academy of Sciences, Beijing, China ruiqi.song@ia.ac.cn, knightdby@gmail.com

<sup>3</sup>Waytous Inc., Beijing, China

<sup>†</sup>Corresponding author.

\*These authors contributed equally to this work.

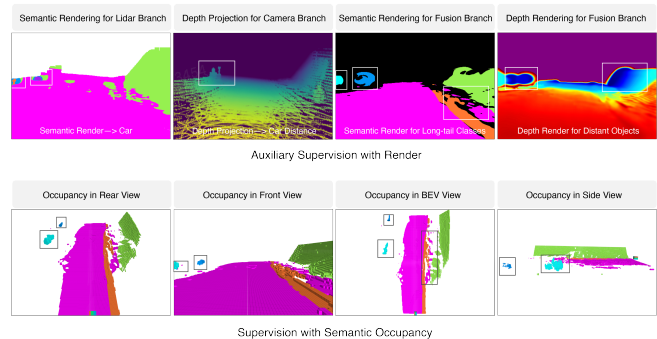


Fig. 1: Semantic and Depth Rendering in Unstructured Scenes with 3D Gaussian Splatting.

Currently, most 3D semantic occupancy prediction methods rely solely on image inputs [3]–[5]. While images provide abundant texture and semantic information, they lack real-world scale and fail to provide absolute depth information. In contrast, LiDAR sensors provide 3D point clouds at centimeter level with precise distance measurements but suffer from relatively low spatial resolution and lack semantic information.

Therefore, fusing image and LiDAR modalities is essential to improve 3D perception performance, especially in complex unstructured scene. Although there have been several studies on multimodal fusion for 3D semantic occupancy prediction [6], [7], effective multi-sensor fusion for unstructured scenes, characterized by sparse geometric features and weak semantic cues, remains underexplored [8]. In addition, the long-tail distribution issue is further exacerbated in unstructured scenes.

To tackle these challenges, we propose UnsOcc, a novel multi-modal fusion framework for 3D semantic occupancy prediction in unstructured scenes. It introduces RenderFusion, a render-based cross-modal alignment mechanism, and GSRefinement, a detail-aware 3D Gaussian Splatting auxiliary supervision module. RenderFusion performs cross-branch rendering to align image and LiDAR features. The image branch uses a depth prediction network supervised by projected LiDAR points, while the LiDAR branch predicts semantic occupancy and renders it into 2D semantics via Gaussian Splatting, supervised by image-based segmentation. This dual-supervision enhances cross-modal feature alignment and fusion. Furthermore, GSRefinement is used to address the severe sparsity of occupancy information in 3D voxel space. During the training phase, it not only predicts 3D semantic occupancy distributions but also projects them into 2D semantic segmentation maps for auxiliary

supervision, with an emphasis on improving recognition of long-tailed classes. This strategy effectively improves the model’s ability to predict long-tailed classes. Figure 1 illustrates a series of rendering results produced by our method. To facilitate the deployment of autonomous driving in unstructured scenes, we developed a n open-pit mine dataset for 3D semantic occupancy prediction and conducted related experiments.

To summarize, our main contributions are as follows: 1) We propose a novel 3D semantic occupancy prediction framework that improves the robustness of prediction in unstructured scenes. 2) Bidirectional supervision for cross-modal feature alignment mechanism and detail-aware 3D Gaussian Splatting auxiliary supervision mechanism are proposed to enhance the capability of cross-modal fusion and long-tail class prediction in unstructured scene, respectively. 3) A dataset specifically designed for autonomous driving in unstructured scene has been constructed, and extensive experiments have been conducted.

## II. RELATED WORK

### A. 3D Semantic Occupancy Prediction

3D occupancy prediction aims to estimate the geometric and semantic occupancy of the surrounding space, which is crucial for autonomous driving and robotics. Early approaches to 3D occupancy prediction primarily relied on monocular depth estimation [3], [9], [10]. MonoScene [3] and OccDepth [9] lifted 2D features into 3D voxel grids through monocular depth estimation. VoxFormer [10] generates reliable queries by predicting image depth, enabling more accurate 3D scene understanding. While these approaches proved effective under limited camera views, they struggled with depth ambiguities, particularly in challenging scenarios. To overcome these limitations, panoramic or surround-view methods were subsequently proposed [4], [11], [12]. Techniques like OccNet [12] and Occ3D [4] leverage multiple camera perspectives to generate more accurate 3D occupancy maps, effectively mitigating depth ambiguities through cross-view information. However, these methods still rely on dense voxel grids, which incur significant computational costs. As a result, recent approaches have begun to explore alternative representations beyond voxels. [13] replaces dense voxel grids with point-based representations, which allow inputs of arbitrary scale and location, increasing modeling flexibility. TPVFormer [14] extends the BEV by adding two additional perpendicular planes to form the Tri-Perspective View (TPV), replacing voxel-based representations and achieving a balance between performance and computational complexity. GaussianFormer [15] utilizes sparse Gaussian primitives to model the scene, achieving accurate scene perception through the optimization of their properties. In this paper, we propose a novel alignment fusion scheme and projection-based supervision to enhance the performance of 3D occupancy prediction in unstructured scenes.

### B. Scene Representation with Rendering

NeRF [16] introduced a neural rendering framework that models a scene as a continuous volumetric function, synthesizing novel views from sparse input images. It has inspired various scene reconstruction techniques [17], [18], with some adaptations for scene perception tasks [19], [20]. However, NeRF-based approaches often incur high computational costs. To overcome these limitations, 3D Gaussian Splatting [21] has been proposed as an efficient and explicit alternative. By using a set of anisotropic Gaussian primitives and employing splat rendering, 3D GS enables real-time rendering. Compared to NeRF, the explicit use of Gaussian primitives allows for a more direct and clear representation of geometric structure and occupancy information, while offering higher efficiency in rendering large-scale scenes. Due to these advantages, existing scene perception methods have also adopted 3D Gaussian Splatting [15], [22]–[25]. GaussianFormer [15] adopts sparse 3D semantic Gaussian primitives to represent the scene. It iteratively optimizes the properties of the Gaussian primitives and employs an efficient Gaussian-to-voxel splatting method to generate 3D occupancy predictions. GaussRender [25] is not confined to projecting 3D voxels onto the camera viewpoint. It introduces the projection of 3D occupancy predictions and ground truth to arbitrary viewpoints for loss computation. Gaussianbev [24] models the scene with a set of 3D Gaussian primitives and leverages 3D Gaussian splatting, in place of LSS method, to acquire BEV representation.

### C. Feature Fusion for 3D Scene Representation

The image and LiDAR modalities exhibit significant complementarity in perception tasks. Therefore, fusing image and LiDAR data can overcome the limitations of each individual modality, offering a more comprehensive and robust scene perception capability. A large number of perception solutions, including 3D object detection [26], [27] and semantic segmentation tasks [28], [29], have already adopted fusion schemes that combine image and LiDAR data. In the context of 3D occupancy prediction tasks, there has also been research focusing on multimodal fusion [6], [7], [30], [31]. OccFusion [7] fuses the mapped image features and point cloud features by concatenating them along the channel dimension and applying 3D convolutions. Co-Occ [30] introduces the K-Nearest Neighbor (KNN) algorithm to optimize the concatenation process, alleviating the issue of some voxels containing only image or point cloud features. CONet [6] and Occgen [31] use 3D convolutions to apply weighted operations on the image and point cloud feature branches and perform fusion through weighted summation, thus improving the effectiveness and robustness of the fusion results. Unlike the fusion schemes mentioned above, we introduce 3D Gaussian rendering to implement a cross-supervision alignment paradigm, which forms the basis for constructing a more effective fusion module.

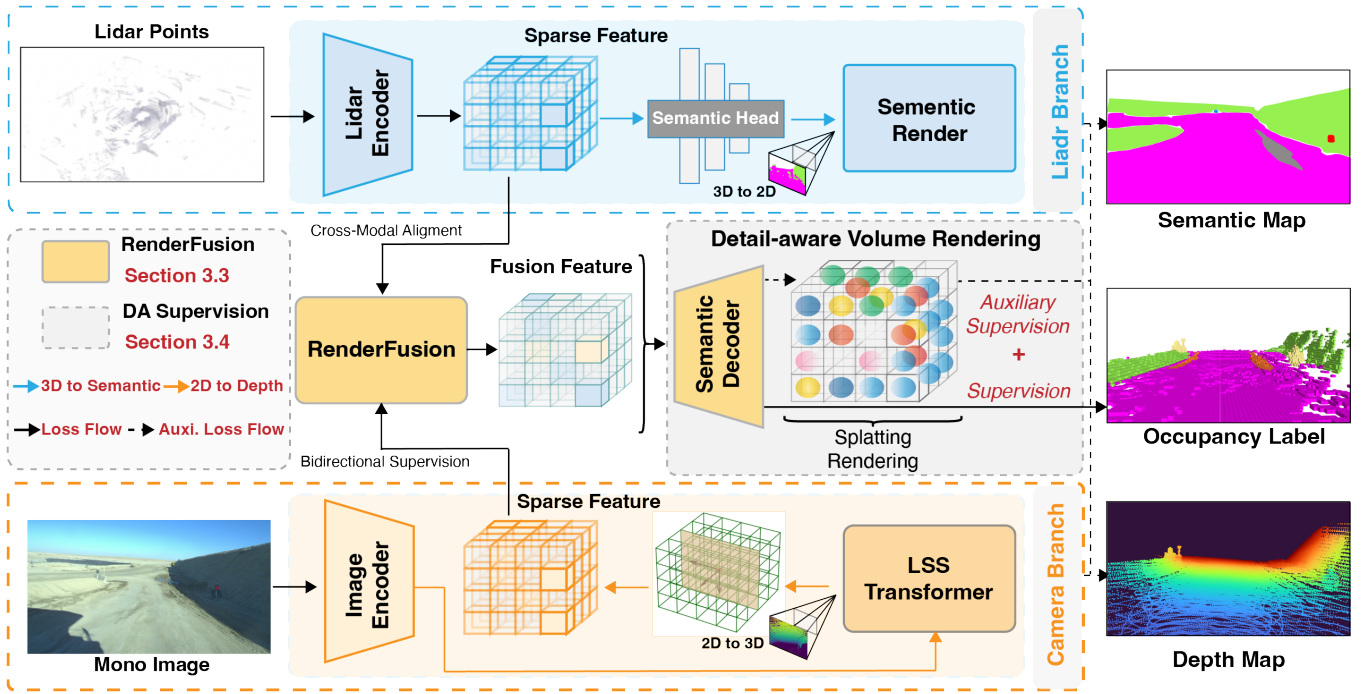


Fig. 2: **Framework of our UnsOcc.** Features from image and LiDAR modalities are extracted, aligned via RenderFusion, and fused. The fused features are used for 3D occupancy prediction, with auxiliary supervision provided by 2D semantic rendering through 3D Gaussian Splatting.

### III. METHOD

#### A. Overview

The overall architecture of our model is illustrated in Figure 2. At its core, we incorporate 3D Gaussian Splatting, a technique that has recently been widely adopted due to its faster rendering speed compared to volume rendering. By representing a scene as a set of 3D Gaussians, each Gaussian distribution  $G$  is defined as

$$G(X) = e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}, \quad (1)$$

where  $\mu$  and  $\Sigma$  denote the mean and 3D covariance matrix, respectively. Through projection, each 3D Gaussian is mapped to a 2D Gaussian  $G_2$ , with its covariance matrix computed as

$$\Sigma' = JW\Sigma W^T J^T, \quad (2)$$

where  $W$  is the view transformation matrix and  $J$  the Jacobian matrix. The rendering process then applies alpha-blending to compute pixel colors:

$$c(x) = \sum_{i=1}^N c^i \alpha^i G_2^i(x) \prod_{j=1}^{i-1} (1 - \alpha^j G_2^j(x)), \quad (3)$$

with  $x$  denoting the pixel position, and  $c^i$ ,  $\alpha^i$  the color and opacity of the  $i$ -th Gaussian.

Building on this rendering mechanism, our framework introduces Bidirectional Supervision for Cross-Modal Feature Alignment and Detail-aware Auxiliary Supervision via Gaussian Splatting. The former enhances feature extraction and alignment across modalities through cross-modal supervision, thereby strengthening their fusion. The latter

leverages 2D ground truth as an auxiliary supervision signal to supplement non-empty semantic samples, effectively addressing the sparsity of map elements in unstructured scenes.

#### B. Feature Extraction

Our framework adopts a dual-branch backbone to process the image and LiDAR modalities in parallel.

For the image branch, we use a ResNet backbone pre-trained on ImageNet to capture hierarchical visual patterns. The multi-level features are aggregated by a SECONDFPN neck, which enhances semantic abstraction while preserving spatial details, and outputs unified 128-dimensional representations suitable for downstream tasks.

For the LiDAR branch, the raw point cloud is voxelized into regular 3D grids, with each voxel retaining up to a fixed number of points. A voxel feature encoder based on mean pooling aggregates local geometry, and a sparse convolutional encoder further extracts high-dimensional descriptors that efficiently encode structural and spatial information.

In this way, the image branch contributes rich semantic cues such as textures and object boundaries, while the LiDAR branch provides precise geometric and depth-aware representations. The complementary features are subsequently integrated in the fusion module to enable robust cross-modal reasoning.

#### C. Bidirectional Supervision for Cross-Modal Feature Alignment.

To address the fusion challenges arising from the lack of semantic information in the LiDAR branch and depth information in the cameras branch, we propose a bidirectional

supervision for cross-modal feature alignment mechanism called RenderFusion as shown in Fig.2. This method introduces semantic supervision to the LiDAR branch and depth supervision to the camera branch through a rendering-based approach.

**Depth Rendering for Camera Branch.** In the image branch, since the monocular image lacks depth information, we utilize the depth projection from the LiDAR branch as a supervisory signal to supervise the depth prediction in the image branch. In this way, depth information that is geometrically aligned with the LiDAR branch is introduced into the image branch, enhancing its spatial understanding and facilitating cross-modal consistency.

Specifically, given an input image  $I \in \mathbb{R}^{H \times W \times 3}$ , feature  $F_I \in \mathbb{R}^{H \times W \times C}$  is first extracted through a backbone and neck network and then passed through a Depth Net [32] to produce a depth distribution  $p_d \in \mathbb{R}^{D \times H \times W} = \Phi_d(F_I)$ .  $D$  represents the number of depth bins.  $F_I$  is lifted into image-depth space by weighted aggregation based on the depth distribution:

$$F_{\text{img-d}}(u, v, d) = p_d(u, v, d) \cdot F_I(u, v). \quad (4)$$

By leveraging intrinsic  $K$  and extrinsic  $E$ , the discrete depth coordinates of each pixel  $(u, v, d)$  can be mapped to corresponding 3D voxel coordinates without learnable parameter:

$$F_{\text{vi}}(i, j, k) = \sum_{T(u,v,d)=(i,j,k)} F_{\text{img-d}}(u, v, d), \quad (5)$$

where  $F_{\text{vi}}$  and  $T$  denote the 3D voxel features of the image and the transformation function that maps image depth coordinates to 3D voxel coordinates, respectively.

The subsequent fusion of voxel-level features is highly dependent on the geometric alignment between the two modalities, whereas the transformation of image features from 2D to 3D at this stage critically relies on the accuracy of depth prediction. To ensure accurate and consistent depth prediction, during training, the point cloud input  $P \in \mathbb{R}^{N \times 4}$  from the LiDAR branch is projected to generate a depth map  $Z_L \in \mathbb{R}^{H \times W}$ . By normalizing  $Z_L$  and applying one-hot encoding, we obtain the ground truth for the depth distribution  $Z_p \in \mathbb{R}^{D \times H \times W}$ :

$$Z_L(u', v') = \min(P_{\text{img}}(u, v)), \quad P_{\text{img}} = K \cdot [R \mid t] \cdot P^h, \quad (6)$$

where  $P^h$  denotes the homogeneous coordinates of a LiDAR point,  $[R \mid t]$  is the extrinsic matrix,  $K$  is the intrinsic matrix,  $P_{\text{img}} \in \mathbb{R}^{D \times H \times W}$  is the projected 2D image coordinate.  $u'$  and  $v'$  are the results of rounding the  $u$  and  $v$  components, which represent the horizontal and vertical components of the image coordinates, respectively.

**Semantic Rendering for Lidar Branch.** In the LiDAR branch, we also aim to leverage information from the image branch to improve the alignment between LiDAR and image features. Semantic prediction from LiDAR signals has been widely validated as effective. To facilitate this, we introduce a semantic prediction head that converts 3D LiDAR features into 3D semantic logits. These logits are then projected onto

the image plane using 3D Gaussian Splatting and supervised with 2D semantic segmentation maps obtained from the input images.

Concretely, let  $P \in \mathbb{R}^{N \times 4}$  be the input point clouds. The voxelized  $P_v \in \mathbb{R}^{M \times 4}$  contains the mean values of  $M$  non-empty voxels. Building upon the voxelized representation, subsequent sparse convolutions hierarchically aggregate local voxel neighborhoods, resulting in sparse LiDAR feature representations  $F_L$ . Prior to multi-modal fusion, we establish semantic consistency between sparse LiDAR features and image features through an auxiliary learning task. During training, the sparse feature  $F_L \in \mathbb{R}^{M \times C}$  is processed by a lightweight segmentation head  $\Phi_{\text{seg}3d}$  to produce volumetric semantic logits  $\mathcal{L}_L \in \mathbb{R}^{X \times Y \times Z \times K}$ .  $K$  is the number of 3D occupancy classes. To leverage the information from the image branch as a supervision signal,  $\mathcal{L}_L^3$  is then rendered to the 2D image plane via alpha-blending:

$$\mathcal{L}_L^2 = \sum_{i=1}^N \mathcal{L}_L \cdot \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (7)$$

where  $N$  represents the number of Gaussians involved in the projection,  $\mathcal{L}_L^2$  is the 2D logits obtained from the projection,  $\alpha'$  is the effective opacity computed from the original opacity and Gaussian density. At this point, we can supervise the LiDAR voxel features  $F_L$  using 2D semantic segmentation  $S_I$  that matches the image input.

After pre-fusion optimization, the enhanced LiDAR and image features are fused in 3D voxel space. The LiDAR and projected image features are first spatially aligned, and for each non-empty LiDAR voxel, neighboring image features are retrieved to provide semantic guidance. These neighboring features are then weighted and combined with the LiDAR features to produce the final fused representation.

#### D. Detail-aware auxiliary supervision via Gaussian splatting

In unstructured scenes, particularly in large-scale open environments or natural settings, some object classes typically occupy only a small fraction of the space, with the majority of the area remaining empty. Consequently, more than 98% of the voxels in the 3D occupancy ground truth are empty, resulting in an extremely low density of meaningful semantic occupancy data. This severe sparsity substantially limits the effectiveness of semantic class supervision. However, in the image space, the proportion of empty classes is significantly reduced. To leverage this advantage, we propose a dense 2D auxiliary supervision mechanism, which employs 3D Gaussian Splatting to exploit semantic information from the image space for 3D occupancy prediction. This strategy substantially increases the number of learning samples for non-empty categories, thereby enhancing the model's ability to accurately perceive these classes, especially for rare classes.

In particular, the fused features  $F_f$  is passed through a decoder and a semantic head to output 3D occupancy logits predictions  $\mathcal{O}$ . Moreover, We employ 3D Gaussian Splatting to project  $\mathcal{O}$  onto the camera plane, yielding 2D segmentation

predictions  $\mathcal{L}_{occ}^2$  and depth predictions  $\mathcal{L}_d$

$$\mathcal{L}_{occ}^2 = \sum_{i=1}^N \mathcal{L}_{occ} \cdot \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (8)$$

$$\mathcal{L}_d^2 = \sum_{i=1}^N d_i \cdot \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (9)$$

where  $\mathcal{L}_{occ}$  represents the occupancy logits produced by the Semantic Decoder, and  $d_i$  denotes the center depth of the  $i$ -th Gaussian (the  $z$ -coordinate in the camera coordinate system).

Building upon the use of 3D occupancy ground truth for supervision, we also use 2D ground truth, which matches the input image, to supervise the predictions, refining the learning of long tail samples.

### E. Loss Function

**Depth Predict Loss.** The binary cross-entropy (BCE) loss is adopted to supervise the discrete depth distribution:

$$\mathbf{L}_d = -[Z_p \log(p_d) + (1 - Z_p) \log(1 - p_d)], \quad (10)$$

where  $p_d$  denotes the predicted depth probability and  $Z_p$  represents the corresponding ground-truth label.

**Depth Render Loss.** The SILog loss calculates the loss between the depth predicted  $Z_r$  by 3D Gaussian Splatting and the ground truth depth  $Z_L$ :

$$\mathbf{L}_{dr} = \frac{1}{N} \sum_{i=1}^N |\log(Z_r(i)) - \log(Z_L(i))|. \quad (11)$$

**Segmentation Render Loss.** Cross-entropy loss is applied to both the 2D logits rendered from LIDAR features  $\mathcal{L}_L^2$  and those rendered from 3D occupancy predictions  $\mathcal{L}_{occ}^2$ :

$$\mathbf{L}_{2D} = -\log(\text{Sigmoid}(\mathcal{L}_L^2)_y) - \log(\text{Sigmoid}(\mathcal{L}_{occ}^2)_y) \quad (12)$$

where  $y$  represents the ground truth class label. With the addition of the 3D occupancy loss  $\mathbf{L}_{occ}$ , the total loss is expressed as follows:

$$\mathbf{L} = \mathbf{L}_{occ} + \mathbf{L}_d + \mathbf{L}_{dr} + \mathbf{L}_{2D}. \quad (13)$$

## IV. RESULTS

### A. Datasets

**Open-pit Mine Dataset.** We evaluate our proposed model on a real-world unstructured scene dataset, collected from multiple open-pit mines in China. The sensing platform is equipped with a front-mounted LiDAR, a monocular camera, and an inertial navigation system (INS). Data were gathered under diverse conditions, including daytime, nighttime, rain, snow, and dusty weather, covering both structured roads and unstructured terrains such as muddy or rugged surfaces. In total, the dataset comprises 135 sequences with an average length of 55 frames, featuring complex terrain, loosely distributed objects, and irregular obstacles. Each frame provides synchronized LiDAR point clouds and monocular images.

For ground-truth construction, dynamic objects were removed from multi-sweep LiDAR data to generate dense

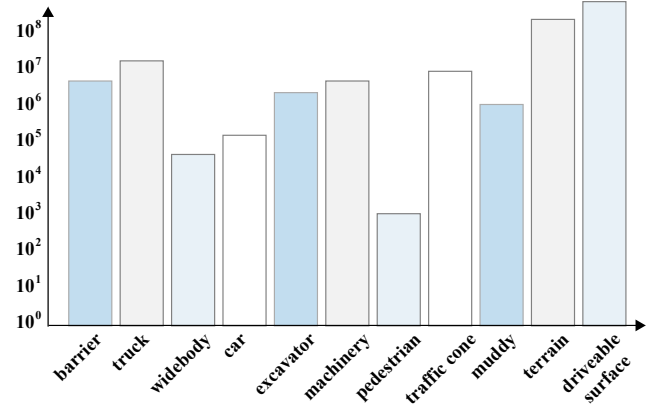


Fig. 3: Distribution of semantic classes in the dataset.

static point cloud maps, which were then voxelized into 3D semantic occupancy labels. The occupancy space is represented as a voxel grid of size  $[256, 256, 32]$ , with a voxel resolution of  $[0.3\text{m}, 0.3\text{m}, 0.3\text{m}]$ . All voxels are categorized into 13 classes, including 11 semantic classes (e.g., driveable surface, terrain, muddy regions, traffic signs, barrier, truck, widebody, car, excavator, machinery and pedestrian), along with an empty and an unknown class. To better illustrate the dataset composition, Fig. 3 shows the voxel distribution across these categories. The semantic voxel distribution is highly sparse, where “driveable surface” and “terrain” dominate the majority while other categories occur rarely, leading to a pronounced long-tail distribution. These sparse categories can be considered as long-tail categories. Furthermore, the dataset contains a variety of irregular obstacles and unstructured elements, which increase the difficulty of perception and make it well-suited for evaluating robust 3D occupancy prediction. Among the 135 sequences, 100 are used for training, 10 for validation, and 25 for testing. Model performance on 3D occupancy prediction is evaluated on the test set using mean Intersection-over-Union (mIoU) and the average IoU over long-tail classes (mIoU(long-tail)).

**nuScenes Dataset.** nuScenes [35] is a large-scale autonomous driving benchmark dataset, consisting of 1,000 driving scenes, of which 700 are used for training, 150 for validation, and 150 for testing. The dataset adopts 3D occupancy labels provided in [4], comprising 17 categories in total, including 16 semantic classes and one empty class. The voxel resolution is set to  $[0.5\text{ m}, 0.5\text{ m}, 0.5\text{ m}]$ , corresponding to a voxel grid of size  $[200, 200, 16]$ .

### B. Evaluation Metrics

We evaluate the 3D occupancy prediction performance of our model using mean Intersection-over-Union (mIoU) and Intersection-over-Union (IoU), defined as follows:

$$\text{IoU} = \frac{TP_{\neq c_0}}{TP_{\neq c_0} + FP_{\neq c_0} + FN_{\neq c_0}} \quad (14)$$

$$\text{mIoU} = \frac{1}{|\mathcal{C}| - 1} \sum_{i \in \mathcal{C}, i \neq c_0} \frac{TP_i}{TP_i + FP_i + FN_i} \quad (15)$$

where  $\mathcal{C}$  denotes the set of all classes, and  $c_0$  represents the empty class. In addition, we also report the mIoU over long-












Method	Modality	driveable surface (73.99%)	terrain (23.70%)	muddy (0.12%)	traffic sign (0.41%)	barrier (0.34%)	truck (1.04%)	widebody (0.007%)	car (0.01%)	excavator (0.14%)	machinery (0.22%)	pedestrian (0.007%)	mIoU	mIoU (long-tail)
														
MonoScene [3]	C	30.11	4.82	0.10	0.62	1.91	5.09	0.00	0.12	0.00	0.41	0.00	3.92	0.92
TPVFormer [14]	C	30.08	5.14	0.15	0.50	4.07	5.15	0.00	0.02	0.00	0.69	0.00	4.17	1.18
Occformer [5]	C	30.93	8.85	11.16	6.95	5.91	8.98	0.00	3.76	0.00	6.91	0.00	4.69	4.85
CGformer [33]	C	<b>61.07</b>	31.05	5.78	16.25	10.92	22.92	0.22	4.69	8.06	11.96	0.00	15.72	8.98
L2COcc(C) [34]	C	60.35	30.99	4.34	18.70	<b>12.07</b>	23.53	0.08	7.69	2.75	15.38	0.00	15.99	9.39
Co-Occ(L) [30]	L	37.26	29.72	7.72	19.47	9.93	31.4	0.00	20.92	12.81	25.98	0.00	17.75	14.25
Co-Occ(C&L) [30]	C&L	37.86	29.68	8.34	<b>21.3</b>	9.43	32.6	0.16	19.21	<b>16.61</b>	28.38	0.00	18.5	15.11
UnsOcc(Ours)	C&L	37.24	29.37	<b>14.70</b>	21.09	8.89	<b>32.17</b>	<b>15.55</b>	<b>23.1</b>	13.63	<b>30.55</b>	0.00	<b>20.57</b>	<b>17.74</b>

TABLE I: Evaluation and comparison on open-pit mine test set.

tail categories by excluding *drivable surface* and *terrain* from the semantic classes in the Opne-pit Mine dataset.

### C. Implementation details

Experiments are conducted on 8 NVIDIA L40 GPUs with a batch size of 1 per GPU. Models use ResNet-50 for the Open-pit Mine dataset and ResNet-101 for nuScenes, trained with AdamW (lr=0.0001, weight decay=0.01) for 24 epochs. Baselines follow their official implementations.

### D. Main Results

**Results on Open-pit Mine Dataset.** Experiments are conducted on the Open-pit Mine Dataset, comparing the proposed approach with various methods with different input modalities, including image-only [3], [5], [14], [33], [34], LiDAR-only [30], and LiDAR-image fusion [30] methods. Table II shows the per-class IoU and overall mIoU. Early image-only methods [3], [5], [14] perform poorly. Compared to recent image-only approaches [33], [34], our method improves mIoU by 30.6% and 28.6%. Against recent LiDAR-only methods [30], our method achieves 15.9% higher mIoU, showing the benefit of multi-modal fusion. Compared to LiDAR-image fusion methods, our approach achieves an 11.19% mIoU gain, with notable improvements on rare classes. It is worth noting that our method achieves better performance on the muddy class compared to all other baselines. The lack of LiDAR returns in muddy regions significantly limits the effectiveness of LiDAR-based methods. Even advanced fusion strategies, such as Co-Occ [30], fail to yield notable gains in these challenging areas. In contrast, our RenderFusion leverages image-guided rendering supervision to compensate for the LiDAR deficiency, leading to notable improvements and demonstrating the effectiveness and robustness of our approach. Moreover, our method achieves a significant improvement in mIoU (long-tail) over all the above methods, demonstrating its superiority in handling

long-tail categories. Figure 4 presents the qualitative comparisons between our method and existing approaches.

**Results on nuScenes Dataset.** To further assess the generalizability of our model, we conduct supplementary experiments on the nuScenes. Our approach achieves higher mIoU compared to several previous state-of-the-art methods, demonstrating its capability to capture both coarse and fine-grained scene structures. Our method shows clear advantages on challenging objects such as traffic cones and pedestrians, indicating stronger sensitivity to fine details. These observations collectively support the robustness of our model and its applicability across different datasets and scenarios.

### E. Ablation Study

**Effect of the GSRefinement.** As shown in Table III, incorporating the Detail-aware module into the baseline leads to a 1.13 improvement in mIoU. Furthermore, the mIoU(long-tail classes) improves by 1.45. This demonstrates that the Detail-aware component effectively enhances the perception capability. The introduction of 2D supervision signals significantly improves semantic prediction performance, especially for rare classes.

**Effect of the RenderFusion.** We further evaluate the effectiveness of the proposed RenderFusion module. As shown in Table III, introducing RenderFusion leads to an additional improvement of 0.55 in mIoU and 0.75 in rare class mIoU. This demonstrates that the proposed multi-modal feature alignment strategy further enhances the quality of fused features and improves recognition performance on both common and rare semantic categories.

**Combination of GSRefinement and RenderFusion.** When both GSRefinement and RenderFusion are introduced, the mIoU and mIoU (long-tail) further improve. The result shows that introducing RenderFusion on top of GSRefinement brings more gains. Conversely, introducing GSRefinement on top of RenderFusion also yields additional improvements.

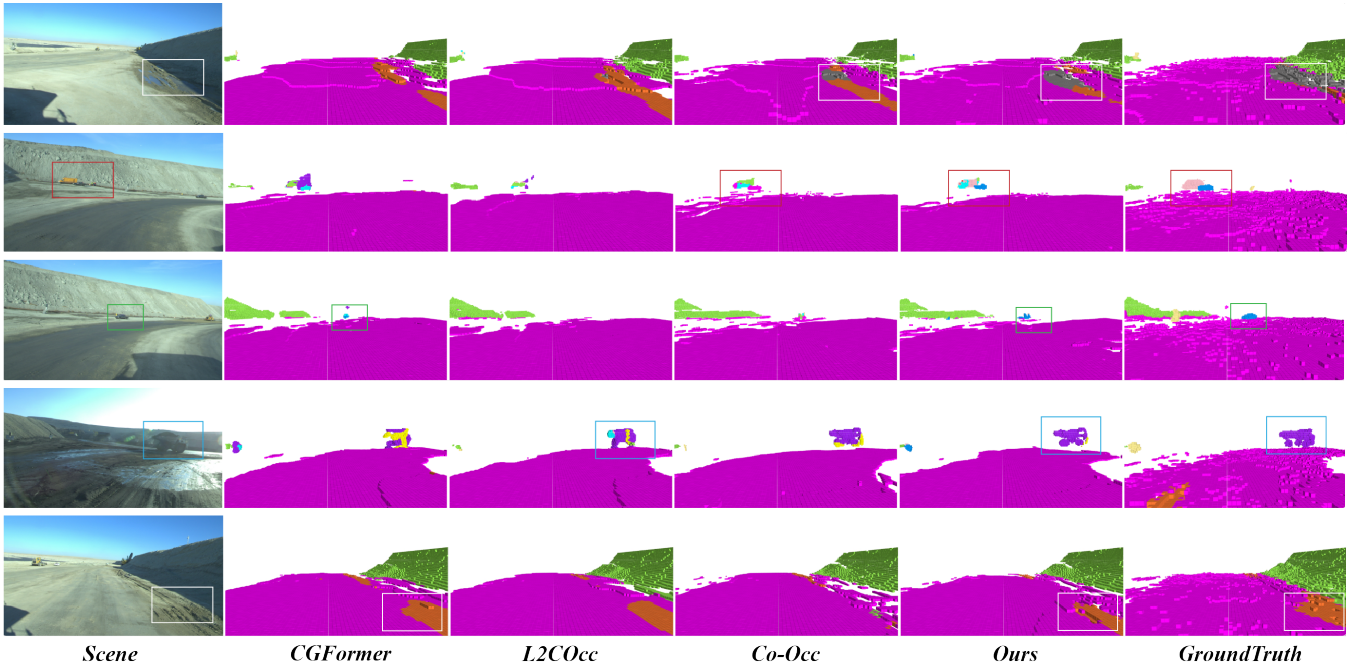


Fig. 4: Comparison of the performance of different 3D semantic occupancy prediction methods.

Method	Modality	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrian	manmade	vegetation	mIoU
		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
MonoScene [3]	C	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	7.3
BEVFormer [36]	C	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	16.7
SurroundOcc [4]	C	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	20.3
OccFormer [5]	C	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	20.1
C-CONet [6]	L	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	18.4
FB-OCc [37]	C	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	19.6
RenderOcc [20]	C	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	19.0
LMSCNet [38]	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	14.9
L-CONet [6]	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	17.7
M-CONet [6]	C&L	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	24.7
Co-Occ [30]	C&L	28.1	16.1	34.0	<b>37.7</b>	17.0	21.6	20.8	15.9	21.9	28.7	42.3	25.4	29.1	<b>28.6</b>	28.2	38.0	27.1
UnsOcc(Ours)	C&L	<b>29.8</b>	<b>18.2</b>	<b>34.5</b>	37.2	<b>18.2</b>	<b>22.1</b>	<b>21.6</b>	<b>22.7</b>	<b>22.0</b>	<b>29.9</b>	<b>42.6</b>	<b>26.1</b>	<b>29.2</b>	28.4	<b>28.6</b>	<b>38.1</b>	<b>28.1</b>

TABLE II: Evaluation and comparison experiments on nuScenes-Occupancy validation set.

GSRefinement	RenderFusion	Resolution	modality	mIoU	mIoU (long-tail)
✗	✗	384×1280	C&L	18.50	15.11
✓	✗	384×1280	C&L	19.63	16.56
✗	✓	384×1280	C&L	19.08	15.86
✓	✓	384×1280	C&L	20.57	17.74

TABLE III: Ablation study on architectural components.

#### F. Deployment

We deployed our proposed method on different vehicle types at multiple mining sites, with vehicles equipped with monocular and multi-view cameras. The deployment results show that the fusion strategy provides significant gains with monocular input and remains effective under multi-view settings, consistent with the aforementioned experimental results, further validating the robustness and practicality of

our approach. Additional qualitative results are available in the supplementary video.

#### V. CONCLUSION

This paper proposes a multimodal 3D occupancy prediction framework that integrates RenderFusion with GSRefinement, leveraging complementary image and LiDAR information to enhance scene understanding. The framework introduces 3D Gaussian Splatting to enable cross-modal supervision between branches, improving feature alignment before voxel-level fusion. It further incorporates semantically rich 2D signals for complementary supervision, boosting occupancy prediction performance. Experiments on the open-pit mine and nuScenes datasets validate the effectiveness of the proposed method.

## REFERENCES

- [1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 12 697–12 705.
- [2] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.
- [3] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 3991–4001.
- [4] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 21 729–21 740.
- [5] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 9433–9443.
- [6] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 17 850–17 859.
- [7] Z. Ming, J. Stephany Berrio, M. Shan, and S. Worrall, "Occfusion: A straightforward and effective multi-sensor fusion framework for 3d occupancy prediction," *arXiv preprint arXiv:2403.00000*, 2024.
- [8] S. Chen, J. Liu, X. Liang, S. Zhang, J. Hyppä, and R. Chen, "A novel calibration method between a camera and a 3d lidar with infrared images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4963–4969.
- [9] R. Miao, W. Liu, M. Chen, Z. Gong, W. Xu, C. Hu, and S. Zhou, "Occdepth: A depth-aware method for 3d semantic scene completion," *arXiv preprint arXiv:2302.13540*, 2023.
- [10] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 9087–9098.
- [11] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.
- [12] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 8406–8415.
- [13] Y. Shi, T. Cheng, Q. Zhang, W. Liu, and X. Wang, "Occupancy as set of points," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 72–87.
- [14] Y. Huang, W. Zheng, Y. Zhang, and J. Zhou, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 9223–9232.
- [15] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 376–393.
- [16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [17] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.
- [18] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 5855–5864.
- [19] C. Zhang, J. Yan, Y. Wei, J. Li, L. Liu, Y. Tang, Y. Duan, and J. Lu, "Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields," *CoRR*, 2023.
- [20] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 404–12 411.
- [21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 139:1–139:14, 2023.
- [22] W. Gan, F. Liu, H. Xu, N. Mo, and N. Yokoya, "Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting," *arXiv preprint arXiv:2408.11447*, 2024.
- [23] H. Jiang, L. Liu, T. Cheng, X. Wang, T. Lin, Z. Su, W. Liu, and X. Wang, "Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding," *arXiv preprint arXiv:2412.13193*, 2024.
- [24] F. Chabot, N. Granger, and G. Lapouge, "Gaussianbev: 3d gaussian representation meets perception models for bev segmentation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 2250–2259.
- [25] L. Chambon, E. Zablocki, A. Boulch, M. Chen, and M. Cord, "Gaussrender: Learning 3d occupancy with gaussian rendering," *arXiv preprint arXiv:2502.05040*, 2025.
- [26] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.
- [27] H. Wang, H. Tang, S. Shi, A. Li, Z. Li, B. Schiele, and L. Wang, "Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 6792–6802.
- [28] Y. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li *et al.*, "Uniseq: A unified multi-modal lidar segmentation network and the openpcseg codebase," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 21 662–21 673.
- [29] J. Li, H. Dai, H. Han, and Y. Ding, "Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 21 694–21 704.
- [30] J. Pan, Z. Wang, and L. Wang, "Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction," *IEEE Robotics and Automation Letters*, 2024.
- [31] G. Wang, Z. Wang, P. Tang, J. Zheng, X. Ren, B. Feng, and C. Ma, "Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 95–112.
- [32] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 194–210.
- [33] Z. Yu, R. Zhang, J. Ying, J. Yu, X. Hu, L. Luo, S.-Y. Cao, and H.-I. Shen, "Context and geometry aware voxel transformer for semantic scene completion," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 1531–1555.
- [34] R. Wang, Y. Ma, Y. Yao, S. Tao, H. Li, Z. Zhu, Y. Liu, and X. Zuo, "L2cocc: Lightweight camera-centric semantic scene completion via distillation of lidar model," *arXiv preprint arXiv:2503.12369*, 2025.
- [35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liang, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [36] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [37] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.
- [38] L. Roldao, R. De Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 111–119.