nuCraft: Crafting High Resolution 3D Semantic Occupancy for Unified 3D Scene Understanding

Benjin Zhu[1], Zhe Wang[2], and Hongsheng Li[1,3,4]

[1] MMLab, The Chinese University of Hong Kong
[2] SenseTime Research
[3] Shanghai AI Laboratory
[4] CPII under InnoHK

Abstract. Existing benchmarks for 3D semantic occupancy prediction in autonomous driving are limited by low resolution (up to [512×512×40] with 0.2m voxel size) and inaccurate annotations, hindering the unification of 3D scene understanding through the occupancy representation. Moreover, previous methods can only generate occupancy predictions at 0.4m resolution or lower, requiring post-upsampling to reach their full resolution (0.2m). The root of these limitations lies in the sparsity, noise, and even errors present in the raw data. In this paper, we overcome these challenges by introducing nuCraft, a high-resolution and accurate semantic occupancy dataset derived from nuScenes. nuCraft offers an 8× increase in resolution ([1024 × 1024 × 80] with voxel size of 0.1m) and more precise semantic annotations compared to previous benchmarks. To address the high memory cost of high-resolution occupancy prediction, we propose VQ-Occ, a novel method that encodes occupancy data into a compact latent feature space using a VQ-VAE. This approach simplifies semantic occupancy prediction into feature simulation in the VQ latent space, making it easier and more memory-efficient. Our method enables direct generation of semantic occupancy fields at high resolution without post-upsampling, facilitating a more unified approach to 3D scene understanding. We validate the superior quality of nuCraft and the effectiveness of VQ-Occ through extensive experiments, demonstrating significant advancements over existing benchmarks and methods.

Keywords: 3D Scene Understanding · Semantic Occupancy Prediction

1 Introduction

Over the past few years, autonomous driving (AD) has witnessed the shift towards unified 3D scene understanding, rather than relying on the conventional approach of dividing AD perception into many sub-tasks (*e.g.*, detection, segmentation) then merge their results for final unification. However, AD scenarios are inherently complex, encompassing a wide variety of elements such as pedestrians, vehicles, drivable areas and buildings. This diversity presents a considerable challenge in developing unified approaches to 3D scene understanding. 3D semantic occupancy demonstrates great potential [16, 27, 28] in unifying various

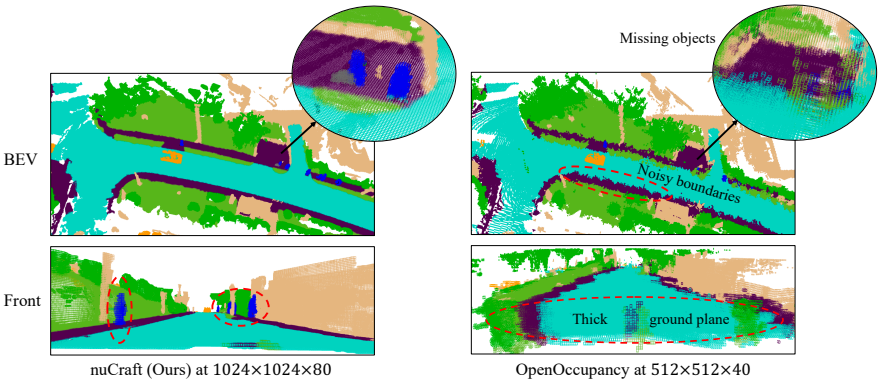

Fig. 1: Occupancy quality comparison. OpenOccupancy (Right) faces problems such as missing objects, noisy road boundaries and thick ground planes. Our nuCraft (Left) can preserve fine details with much less noises at a 8× higher resolution.

sub-tasks and achieving comprehensive 3D scene understanding. This representation involves rasterizing 3D scenes into dense voxel grids, capturing geometric information through voxel occupancy, and semantic information through categorical labels for occupied voxels. The data generation of 3D semantic occupancy in research communities typically involves two stages: a LiDAR aggregation stage that combines multiple LiDAR frames of a sequence to create relatively dense point clouds as reference for initial semantic occupancy ground truth (GT), followed by a "voxel densification" [27] stage that infers and fills empty locations and assigns semantic labels to create denser occupancy. All available GT annotations from existing sub-tasks (*e.g.*, 3D bounding boxes, point semantic labels) can be properly integrated into 3D semantic occupancy. Moreover, the generation of 3D semantic occupancy typically aggregates all available data frames, embedding future frame information into the current frame's occupancy GT, inherently requiring the model to use past information for future state prediction.

Despite its potential, existing benchmarks (mainly derived from widely-used nuScenes [2] dataset) face limitations of low resolution and inaccurate annotations (Fig. 1 right). Due to the sparsity of LiDAR scans, the occupancy resolutions of Occ3D-nuScenes and OpenOccupancy are capped at 0.4m and 0.2m, respectively, whereas conventional voxel-based 3D object detectors typically require input data with 0.1m resolutions or finer. This low resolution compromises the goal of using 3D semantic occupancy to achieve unified 3D scene understanding [28], resulting in significant information loss. Besides, existing occupancy datasets neglect to deal with noises and errors in the raw data (*e.g.*, inaccurate ego poses and wrong annotations, see Supplementary Materials), which greatly harm the results of LiDAR frames aggregation (thick ground plane in Fig. 1) and the inference of semantic labels for non-key frames (noisy road boundaries in Fig. 1) for voxel densification. Consequently, previous benchmarks cannot generate precise enough 3D semantic occupancy GT for models to generate reli-

able semantic occupancy predictions. Furthermore, existing methods [31,41] still struggle to handle 3D occupancy at 0.2m resolution directly due to the quadratic increase in memory consumption and the increased difficulty in learning geometric relations and semantics in larger occupancy fields. The way they predicting coarse occupancy (*e.g.*, 0.4m) and then upsampling to the full resolution (0.2m) negates the intended purpose of high-resolution occupancy.

In this paper, we address the issue of insufficient resolution and inaccurate annotations in existing 3D occupancy datasets [27, 28, 31] by introducing our high-resolution and precise 3D semantic occupancy dataset, **nuCraft**. By employing pose estimation to correct inaccurate ego poses within driving sequences, we achieve well-aligned aggregated LiDAR point clouds for providing reliable and clean input for subsequent processing. For voxel densification, we utilize a multi-level octree representation of the dense point cloud to maintain a controllable error rate during mesh reconstruction. Additionally, we enhance the consistency of boundaries and reduce noise in semantic labels by incorporating a mesh vertices semantic prediction branch into mesh reconstruction. This approach allows us to generate semantic scene meshes with clear boundaries for occupancy GT generation. As shown in Fig. 1 left, our nuCraft dataset can not only provide 3D semantic occupancy GT with 0.1m resolution ($1024 \times 1024 \times 1024$), but also yields more precise GT (*e.g.*, accurate boundaries, complete object shapes). To address the challenges associated with the increased memory consumption and learning difficulty of high-resolution occupancy, we propose VQ-Occ, a novel method that encodes the high-resolution occupancy GT into a compact latent feature space using Vector Quantized-Variational AutoEncoder (VQ-VAE) [29]. This method enables us to construct a discrete codebook for the entire occupancy GT dataset, transforming the task of 3D semantic occupancy prediction into feature simulation in the VQ feature space. Consequently, the model only needs to simulate the discrete latent features of occupancy GT in a compressed latent space, making the process both easier and more memory-efficient. During inference, the inputs are encoded and projected onto the VQ latent space, then passed to the pre-trained VQ-VAE decoder to obtain final occupancy predictions. VQ-Occ surpasses all previous methods and can directly generate occupancy predictions at high resolutions without upsampling on both OpenOccupancy and our nuCraft, paving the way for unifying 3D scene understanding through 3D semantic occupancy prediction. Contributions of this work are listed as follows:

- We introduce nuCraft, a high-resolution 3D semantic occupancy dataset with 8× resolution and more precise annotations than previous benchmarks.
- We present a general and robust data generation pipeline for creating high-quality 3D semantic occupancy GT from noisy data using only off-the-shelf annotations from 3D detection and segmentation tasks with no human effort.
- We propose VQ-Occ, a novel 3D occupancy prediction framework. It decouples the encoding of occupancy GT and semantic occupancy prediction, achieving direct high-resolution prediction without post-upsampling and better performance than previous methods on OpenOccupancy and our nuCraft.

2 Related Works

2.1 3D Semantic Occupancy Benchmarks

The task of semantic occupancy perception, first introduced by the SUNCG dataset [26], requires algorithms to output both occupancy and semantic labels for all voxels within the camera-view frustum. Several indoor benchmarks have been developed, including [6,9,11,25,33], focusing on stationary indoor environments. However, the availability of datasets for outdoor scenarios is more limited. SemanticKITTI [1] is a pioneering 3D semantic occupancy benchmarks for AD, providing occupancy GT of 0.2m resolution by accumulating LiDAR scans in a sequence with its precise ego poses and point semantic labels. However, it lacks diversity in urban scenes and only evaluates front-view occupancy predictions, limiting the generalization of occupancy perception algorithms. Occ3D [27], OpenOcc [28], and OpenOccupancy [31] provide diverse occupancy scenarios by deriving from the nuScenes dataset series [2,10], but they suffer from limited resolution (0.4m for Occ3D and OpenOcc, and 0.2m for OpenOccupancy), inaccurate boundaries, and missing annotations due to noises and inconsistencies in data processing.

2.2 3D Semantic Occupancy Prediction Methods

3D occupancy prediction is a challenging task in autonomous driving perception, aiming to achieve unified 3D scene understanding by simultaneously predicting occupancy status and semantic labels for all voxels in the surrounding space [5,27,32]. Voxel-based approaches like Voxformer [19] and Occ3D [27] utilize 2.5D information and coarse-to-fine voxel encoders to construct occupancy representations, while RenderOcc [22] extracts 3D volume features from surround views and predicts density and labels for each voxel with NeRF supervision. SelfOcc [15] explore a self-supervised way to learn 3D occupancy using only video sequences. However, voxel-based representations pose challenges in computational complexity. BEV-based methods like FlashOcc [37] represents features on a BEV grid, reducing feature representation in the height dimension. [31,41] report their results at a higher resolution (*i.e.*, 0.2m) but require upsampling from low resolutions to their full resolution. Moreover, the computation cost of occupancy prediction increases quadratically as GT resolution increases, and the learning difficulty also increases as scene geometries become more complex with higher occupancy resolution. Our VQ-Occ deal with these challenges by decoupling semantic occupancy prediction into occupancy GT encoding and feature simulation in the VQ space. It greatly eases the difficulty in learning semantics and geometries at high resolutions, and achieves direct high-resolution occupancy prediction with better performance and no post-upsampling.

2.3 3D VQ-VAE

Vector Quantized-Variational AutoEncoders (VQ-VAEs) [29] have shown impressive results in compressing and representing high-dimensional data in a

compact latent space and generating high-quality new data samples, such as images [7, 23] and point clouds [8, 34]. UltraLiDAR [34] utilizes VQ-VAE to tokenize point clouds and uses discrete tokens for LiDAR simulation in driving scenarios. [18, 20] extend discrete diffusion models to learn categorical distributions or shape priors on ShapeNet [4]. OccWorld [38] utilize VQ to obtain discrete scene tokens for generative predictive training to model the evolution of the driving scenes. Inspired by the success of VQ-VAEs on point clouds, we leverages VQ to encode the high-resolution semantic occupancy data into a compact latent feature space, followed by feature simulation in the latent space for 3D occupancy prediction.

3 The nuCraft Dataset

3.1 3D Semantic Occupancy Prediction

The objective of 3D semantic occupancy prediction is to determine the state of each voxel in a 3D scene given a sequence of sensor inputs. These inputs can vary in modalities (*e.g.*, sequences of N surround-view camera images with known intrinsic and extrinsic parameters, sequences of LiDAR scans, or a combination of LiDAR, multi-view camera, and other sensor measurements). The GT for this task consists of the states of the voxels, encompassing both occupancy states (such as "occupied" or "empty") and semantic labels. For instance, a voxel occupied by a vehicle would be labeled as ("occupied", "vehicle"), while a voxel in free space would be labeled as ("empty", None). During evaluation, pre-computed sensor visibility masks are used to filter out unobservable voxels (*e.g.*, voxels behind a wall) to conduct evaluation on visible voxels only. We mainly calculate the Intersection of Union (IoU) as the geometric metric to measures whether each voxel is being occupied or empty, as well as mean IoU (mIoU) of each class as the semantic metric.

3.2 Data Generation

The data generation pipeline of nuCraft consists of 3 steps. A pre-processing step to provide better inputs for next steps (*e.g.*, generate longer sequences). A data aggregation step to conduct pose estimation before aggregating lidar scans within sequences. Then a mesh-reconstruction step is applied to generate high-quality semantic meshes for the "voxel densification" purpose. At last, necessary post-processings are adopted to filter out outliers, reduce noises and generate sensor visibility masks for evaluation.

Pre-processing. *Static/Moving Parts Separation.* Objects can be broadly categorized into static and moving types. Separating these two types of objects allows us to consider their unique properties, like scale and rigidity. To do so, we employ a powerful 3D detector CMT [35] to estimate the properties of objects (*e.g.*, velocity of objects) in non-key frames. Additionally, we use MapMOS [21]

to filter moving noise points left in the background. This helps to create a cleaner static background and more complete shapes for moving objects, ensuring the high input quality for next-step densification.
Continuous Scenes Grouping. Scenes in nuScenes are randomly select from continuous driving logs. Thus we merge potentially continuous scenes within a log to form longer sequences. This strategy can utilize more LiDAR scans and generate more complete scene point clouds. Specifically, we identify the starting and ending points of each scene and link continuous scenes based on geometric distances. For example, 750 training scenes can be grouped into ∼350 longer sequences.

LiDAR Sequence Aggregation with Pose Estimation. Considering the sparsity of nuScenes lidar scans and missing z-value in ego pose, we pinpoint that pose estimation is required to realize better frame alignment for LiDAR aggregation. We employ Kiss-ICP [30] for pose estimation, which aligns point clouds from different time frames by minimizing the discrepancy between corresponding points. This step improves consistency of consecutive LiDAR scans, thus provide more reliable point clouds for mesh reconstruction. To infer point semantic labels in non-key frames, we utilize K-Nearest Neighbors (KNN) for propagating semantic labels from key frames to unlabeled sweeps.

Voxel Densification with Semantic Mesh Reconstruction from Octree Representation. Aggregated point clouds still cannot generate occupancy GT in high resolutions due to the sparsity of point cloud. We choose mesh reconstruction to generate more complete 3D scenes. To maintain controllable error rates during reconstruction, we resort to multi-level octree to store point clouds. This structure enables us to limit the reconstruction process to be within leaf nodes, thus limit the reconstruction error and guarantee its correspondence to input point clouds (see Supplementaries Materials). Here we use SHINE-Mapping [39] for its effectiveness in mesh reconstruction with the desired octree properties. By balancing the levels and node resolution of the octree, we can obtain precise scene meshes with limited noises and error. However, the semantic labels for the reconstructed mesh are still unavailable. Considering the noises and wrong semantic labels caused by label propagation in the aggregated lidar point clouds, we incorporate a mesh vertices semantic label prediction branch into SHINE-Mapping, which can learn from noisy labels and generate smooth mesh boundaries with less semantic error. The resulting semantic meshes allow for occupancy data generation with sufficient resolution while significantly reducing noises observed in previous benchmarks like OpenOccupancy (Fig. 1 right). As shown in Fig 1 left, we can generate occupancy GT with less noise and smooth semantic boundaries. For the foreground objects, we employ NKSR [12] to generate more complete object shapes for rigid objects like cars. At last, We employ mesh cleaning techniques to remove outliers, reduces noises and fill empty holes.

Occupancy GT Generation. The final occupancy GT can be obtained by sampling dense and uniform point clouds from the mesh surfaces followed by

Table 1: Comparison of 3D object detection performance.

Method	mAP
OpenOccupancy	58.4
nuCraft@0.2	62.7
nuCraft@0.1	77.9

Table 2: Comparison of drivable area segmentation performance.

Method	IoU	mIoU
OpenOccupancy	71.2	64.1
nuCraft@0.2	78.3	69.8
nuCraft@0.1	84.2	75.7

voxelization with certain voxel size. The semantic labels for each point are derived from its closest vertex. In this way, we can generate 3D semantic occupancy GT at a high resolution (*e.g*, 0.1m or finer). Apart from the occupancy and semantic labels, we also assign available attributes for objects (*e.g.*, velocity) to its corresponding occupancy. In this way, our nuCraft can also be used for occupancy motion prediction. For the visibility mask generation of different sensors, we generally adopt the approach used by Occ3D [27] to generate visibility masks for cameras and LiDAR.

3.3 Evaluation of nuCraft

To evaluate the quality of our nuCraft dataset, we assess its capability to store information from nuScenes' detection and segmentation annotations. Then we ablate design choices of the data generation pipeline.

Foreground: 3D Object Detection We evaluate nuCraft's ability of retaining foreground objects using 3D object detector Voxel-DETR [40]. The occupancy centroids and their semantic labels from nuCraft forms the input (*i.e.*, $N \times 4$ point clouds, N is voxel numbers). The nuScenes mean Average Precision (mAP) metric is used to measure 3D object detection performance. As shown in Table 1, nuCraft enables significantly higher mAP compared to OpenOccupancy, especially at the finer 0.1m resolution. This demonstrates the significance of high-resolution inputs, and nuCraft's ability to effectively preserve geometric and semantic information of foreground objects.

Background: Drivable Area Segmentation To assess nuCraft's accuracy in representing background regions, we construct GT masks for sidewalks and drivable areas using the HD maps of nuScenes. We then compute the 2D IoU and mIoU between the HD map GT masks and the bird's eye view (BEV) projections of the occupancy voxels related to these categories. Table 2 shows that nuCraft achieves higher IoU and mIoU scores compared to OpenOccupancy at all resolutions, indicating more precise representation of drivable areas and background. Moreover, the improved mIoU scores and qualitative visualizations (Fig. 1) demonstrate that nuCraft contains less noise and more accurate annotations compared to previous datasets (See Supplementary Materials for more visualization). The large gains at higher resolutions further underscore the value

Table 3: Component of data generation. The first row denotes the default nuCraft GT at 0.2m resolution. Components are removed one-by-one from the first-row.

Component Removed	mAP	mIoU
None (Full nuCraft@0.2)	**62.7**	**69.8**
S/MPS	62.1	67.2
CSG	62.4	66.5
PE	60.6	66.1
Mesh	58.7	63.2
Clean	62.7	69.7

of nuCraft's high-resolution and precise occupancy in enabling more detailed 3D scene understanding.

Ablations of Data Generation To understand the impact of various components in the nuCraft data generation pipeline, we conduct ablation experiments by removing each component: Static/Moving Parts Separation (S/MPS), Continuous Scenes Grouping (CSG), Pose Estimation (PE), Mesh reconstruction (Mesh), and Mesh cleaning (Clean). Table 3 shows the effect on detection mAP and segmentation mIoU. The results indicate that each component contributes meaningfully to the overall quality of nuCraft. Removing the mesh reconstruction step leads to the largest drop in both mAP (-4.0) and mIoU (-6.6), highlighting its importance in generating high-quality occupancy GT. Pose estimation also has a significant impact, with its removal causing a decrease of 2.1 in mAP and 3.7 in mIoU. The S/MPS and CSG components provide modest improvements of 0.6 and 0.3 in mAP, respectively, and 2.6 and 3.3 in mIoU. The mesh cleaning step appears to have minimal effect on the metrics, suggesting that the mesh reconstruction process already produces high-quality outputs.

The evaluation demonstrates that nuCraft is a high-quality dataset that effectively unifies foreground and background information while providing less noise and more accurate annotations. Our observations not only provide insights for evaluating our nuCraft but also serve as guidelines for creating new and high-resolution occupancy datasets for 3D scene understanding.

3.4 More Occupancy GT Examples of nuCraft

Figure 2 presents additional examples of the occupancy GT from the nuCraft dataset at 0.1m resolution. These visualizations showcase the high level of detail and precision captured by nuCraft, enabling more accurate modeling of complex urban environments for 3D scene understanding tasks.

4 VQ-Occ: Vector Quantized 3D Occupancy Prediction

Previous methods for semantic occupancy perception face significant limitations in processing high-resolution occupancy data. For instance, [31, 41] can only

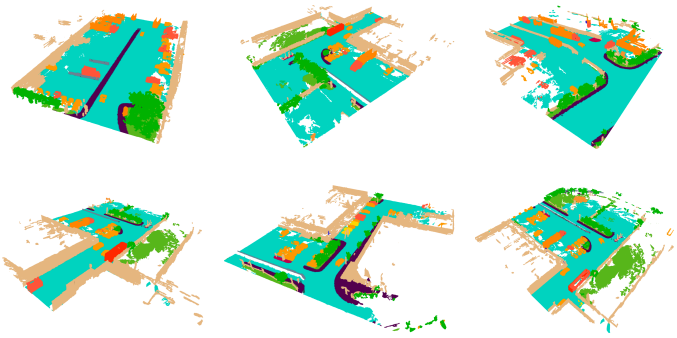

Fig. 2: More occupancy GT examples from our nuCraft dataset with 0.1m resolution.

generate occupancy prediction with 0.4m resolution, followed by upsampling to the full resolution (0.2m on OpenOccupancy). This limitation leads to a loss of fine-grained details crucial for accurate 3D scene understanding. To deal with these challenges, we introduce VQ-Occ (Fig. 3), a novel approach that leverages Vector Quantized-Variational AutoEncoders (VQ-VAE) [29] to efficiently encode high-resolution semantic occupancy GT into a compact latent feature space. By transforming the task of 3D occupancy prediction into discrete feature simulation in the VQ latent space, VQ-Occ achieves better occupancy prediction performance, and achieves direct prediction of 3D semantic occupancy at high resolution (0.2 or finer) without post-upsampling.

4.1 VQ-VAE for 3D Semantic Occupancy Encoding

The encoding of 3D occupancy GT in VQ-Occ is a 3D VQ-VAE similar as [34] that learns to encode and reconstruct the high-resolution occupancy GT. As shown in Fig. 3, the encoder E compresses the occupancy GT $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ into a low-dimensional latent representation $\mathbf{z}_e \in \mathbb{R}^{h \times w \times d \times c}$, where H, W, D are the spatial dimensions, C is the number of semantic classes, and h, w, d, c are the dimensions of the latent space. The latent features are then quantized using a learned codebook $\mathcal{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_K] \in \mathbb{R}^c$ containing K discrete codes. Each code $\mathbf{z}_k \in \mathbb{R}^c$ corresponds to a representative local occupancy pattern. The quantization is performed using a nearestneighbor lookup:

$$\mathbf{z}_q = \mathbf{z}_k, \quad k = \text{argmin}_j \|\mathbf{z}_e - \mathbf{z}_j\|_2^2.$$

The decoder D then reconstructs the occupancy $\hat{\mathbf{x}}$ from the quantized latents $\mathbf{z}_q$. The VQ-VAE is trained using the following objective:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \|\text{sg}[\mathbf{z}_e] - \mathbf{z}_q\|_2^2 + \beta \|\text{sg}[\mathbf{z}_q] - \mathbf{z}_e\|_2^2,$$

where sg denotes the stop-gradient operation, and β is a hyperparameter controlling the commitment loss. By encoding the occupancy GT into a discrete latent space, VQ-Occ can efficiently represent complex 3D scenes using a compact set of learned codes.

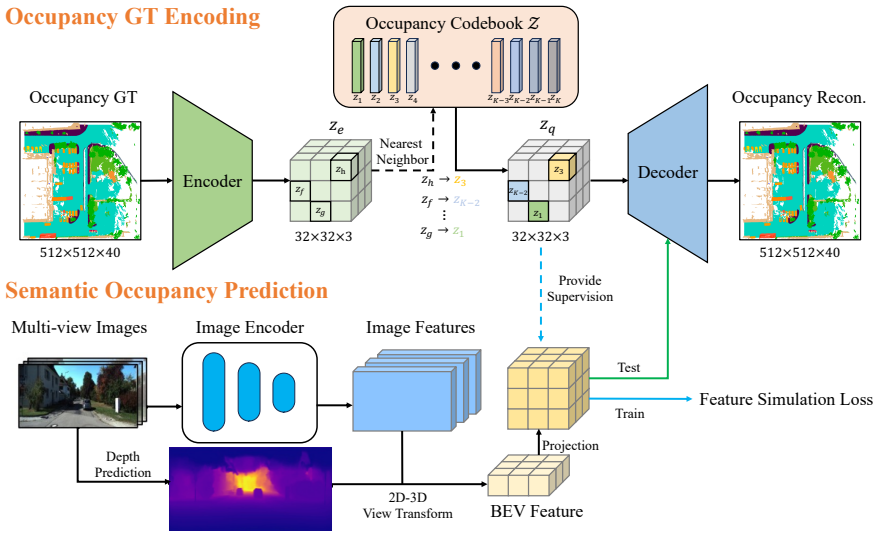

Fig. 3: Pipeline of VQ-Occ. The framework consists of two main components: occupancy GT encoding and semantic occupancy prediction. In the encoding stage, a VQ-VAE is used to compress the high-resolution occupancy GT into a compact latent space using a learned codebook. For occupancy prediction, multi-view images (Here we only visualize image inputs, while LiDAR inputs can be encoded by point cloud backbones and easily integrated to generate the BEV feature.) are encoded by an image encoder to extract features, which are then projected to the VQ latent space dimensions. The model is trained to simulate the discrete VQ features of the corresponding occupancy GT, with an auxiliary depth prediction task to better capture scene geometry. During inference, the image features are encoded, projected to the VQ space, and then decoded by the pre-trained VQ-VAE decoder to directly generate the final high-resolution semantic occupancy predictions without any post-upsampling.

4.2 Occupancy Prediction via Feature Simulation in VQ space

With the pre-trained VQ-VAE, we transform the occupancy prediction task into a feature simulation problem in the VQ latent space. Given the multi-view images or LiDAR frames of a scene, we first extract image features $\mathbf{f}$ using a convolutional backbone from BEVDet [13,14]. The features are then pooled and projected to the dimensions as the VQ latent space:

$$\mathbf{z}_p = P(\text{pool}(\mathbf{f})),$$

where P is a learnable projection head. We train the image encoder to predict latent features $\mathbf{z}_p$ that mimic the pre-computed VQ latents $\mathbf{z}_q$ of the corresponding occupancy GT. The training objective is:

$$\mathcal{L}\text{mimic} = \|\mathbf{z}_p - \text{sg}[\mathbf{z}_q]\|_2^2.$$

Following BEVDet, we also incorporate an auxiliary depth prediction task to encourage the predicted latents to better capture the scene geometry. Different

from previous methods that use LiDAR points to serve as sparse depth supervision, we use depth computed from our high-resolution occupancy GT (0.1m) as supervision. The final training objective is a weighted sum of the feature mimicking and depth losses.

During inference, the image encoder predicts the latent features $\mathbf{z}_p$. These latent features are then quantized by finding the nearest codes in the learned codebook $\mathcal{Z}$ with minial L2 distance between $\mathbf{z}_p$ and each code in the codebook. The quantized latent codes $\mathbf{z}_q$ are then passed through the pre-trained VQ-VAE decoder to generate the full-resolution semantic occupancy $\hat{\mathbf{x}}$. By learning to mimic the compact VQ representation during training, VQ-Occ can efficiently predict high-resolution occupancy without the need for post-upsampling.

5 Experiments

In this section, we present extensive experiments to evaluate the performance of VQ-Occ and demonstrate the effectiveness of our nuCraft dataset. We compare VQ-Occ with state-of-the-art methods on both the OpenOccupancy benchmark and nuCraft. Additionally, we report the reconstruction performance of our VQ-VAE and provide detailed analyses and ablation studies to validate the design choices of VQ-Occ.

5.1 Experimental Setup

Datasets. We train and evaluate VQ-Occ on two datasets: OpenOccupancy and nuCraft. OpenOccupancy, the highest resolution 3D semantic occupancy dataset derived from nuScenes prior to nuCraft, has a resolution of 0.2m. Both datasets are split into training, validation, and test sets following official protocols and contain 16 semantic classes, along with an additional noise/general object class. Results are reported on the 16 semantic classes. The perceptive range extends from $[-51.2\text{m}, -51.2\text{m}, -5\text{m}]$ to $[51.2\text{m}, 51.2\text{m}, 3\text{m}]$, resulting in a volume of $[512 \times 512 \times 40]$ for 3D occupancy prediction. For nuCraft, we primarily evaluate on a 0.2m resolution with the same range and report preliminary results on a 0.1m resolution with direct full-resolution prediction.

Evaluation Metrics. The primary evaluation metric for semantic occupancy prediction is the mean Intersection over Union (mIoU), computed by averaging the IoU of each semantic class. We also report the overall IoU to measure the geometric accuracy of the predicted occupancy.

Implementation Details. The 3D VQ-VAE structure is similar to that of UltraLiDAR [34]. The image encoder adopts the same backbone as BEVDet [13, 14]. We set the codebook size to 2048 with 256 feature dimension, and the latent space dimension to $32 \times 32 \times 3$. For the BEV image features projected in BEV, we add an extra projection layer to elevate features from $32 \times 32 \times 1$ to the same channel as the learned VQ-VAE. We use a voxel size of 0.2m ($[512 \times 512 \times 40]$) unless specified to match the resolution of previous datasets. During training, we employ data augmentation techniques such as random flipping and rotation to avoid overfitting. All experiments are conducted on 8×NVIDIA A100 GPUs.

Table 4: 3D Semantic occupancy prediction results on OpenOccupancy *val* set. VQ-Occ achieves better performance than all previous methods from all input modalities.

Method	Input	IoU	mIoU	barrier	bicycle	bus	car	constr. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	vegetation
MonoScene [3]	C	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3
TPVFormer [16]	C	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0
C-CONet [31]	C	20.1	12.8	13.2	8.1	15.4	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8
VQ-Occ (Ours)	C	**21.5**	**13.6**	**14.1**	**8.8**	**16.4**	**18.3**	**6.8**	**11.9**	**10.7**	**8.9**	**5.1**	**12.9**	**33.2**	**20.0**
LMSCNet [24]	L	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3
JS3C-Net [36]	L	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3
L-CONet [31]	L	30.9	15.8	17.5	5.2	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5
PointOcc [41]	L	34.1	23.9	24.9	19.0	20.9	25.7	13.4	25.6	30.6	17.9	16.7	21.2	36.5	25.6
VQ-Occ (Ours)	L	**35.3**	**24.8**	**25.8**	**19.8**	**21.8**	**26.7**	**13.9**	**26.6**	**31.7**	**18.7**	**17.4**	**22.1**	**37.8**	**26.5**
M-CONet [31]	C&L	29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0
VQ-Occ (Ours)	C&L	**36.8**	**25.5**	**26.5**	**20.5**	**22.6**	**27.7**	**14.4**	**27.6**	**32.7**	**19.4**	**18.1**	**22.9**	**39.0**	**27.5**

Table 5: Semantic occupancy prediction results on nuCraft *val* set at 0.2m resolution.

Method	Input	IoU	mIoU	barrier	bicycle	bus	car	constr. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	vegetation
C-CONet [31]	C	20.8	13.4	14.3	9.1	16.5	18.3	7.4	12.3	11.1	9.4	5.8	13.2	32.5	19.9
VQ-Occ (Ours)	C	**21.9**	**14.1**	**15.2**	**9.9**	**17.5**	**19.4**	**8.0**	**13.0**	**11.8**	**10.0**	**6.2**	**14.0**	**34.3**	**21.1**
L-CONet [31]	L	31.3	16.5	18.6	6.3	14.4	19.2	8.9	6.5	10.7	6.7	14.3	14.7	36.0	22.6
PointOcc [41]	L	34.8	24.6	26.0	20.1	22.0	26.8	14.5	27.1	32.1	19.0	17.8	22.3	37.6	26.7
VQ-Occ (Ours)	L	**36.1**	**25.5**	**26.9**	**20.9**	**22.9**	**27.8**	**15.0**	**28.1**	**33.2**	**19.8**	**18.5**	**23.2**	**38.9**	**27.6**
M-CONet [31]	C+L	29.9	20.7	24.4	14.4	22.3	25.4	16.4	17.0	19.1	14.4	16.4	21.8	34.3	22.1
VQ-Occ (Ours)	C+L	**37.5**	**26.2**	**27.6**	**21.6**	**23.7**	**28.8**	**15.5**	**29.2**	**34.3**	**20.5**	**19.2**	**24.0**	**40.1**	**28.6**

5.2 Results on OpenOccupancy

Table 4 presents the 3D semantic occupancy prediction results on the OpenOccupancy dataset, where our VQ-Occ method outperforms all previous methods across all input modalities (C for camera, L for LiDAR, and C&L for multi-modal inputs) in terms of IoU and mIoU. In the camera-based input category, VQ-Occ demonstrates its ability to accurately capture small and dynamic objects such as bicycle, motorcycle, and pedestrian with an IoU of 21.5% and an mIoU of 13.6%. For LiDAR-based input, VQ-Occ shows significant improvements in handling sparse and irregular LiDAR points, especially in the motorcycle and pedestrian categories, with an IoU of 35.3% and an mIoU of 24.8%. In the multi-modal input setting, VQ-Occ achieves an IoU of 36.8% and an mIoU of 25.5%, indicating its superior ability to effectively fuse camera and LiDAR data for a comprehensive understanding of the 3D scene. Overall, the results demonstrate the effectiveness of VQ-Occ in 3D semantic occupancy prediction, with notable improvements in both geometric accuracy and semantic understanding across different input modalities, indicating its robustness in handling diverse objects and scenarios in the complex urban environment of the OpenOccupancy dataset.

5.3 Results on nuCraft

We compare the performance of VQ-Occ on our nuCraft dataset at 0.2m resolution with other methods in Table 5. VQ-Occ consistently outperforms all baselines across all input modalities. With C+L input, VQ-Occ achieves an IoU of 37.5% and an mIoU of 26.2%, surpassing M-CONet by 7.6% and 5.5%, respectively. The results validate the effectiveness of our nuCraft dataset in providing

Table 6: 3D occupancy prediction results on nuCraft@0.1 with batch size 1.

Method	IoU	mIoU	Mem
VQ-Occ (C)	14.3	9.6	77 GB
C-CONet (0.4m ↑)	6.7	3.8	18 GB
C-CONet (0.1m est.)	N/A	N/A	∼500 GB

Table 7: Cross-validation of VQ-Occ at 0.2m resolution. OO stands for OpenOccupancy, while nC denotes nuCraft.

Setting	IoU	mIoU
OO→nC	18.1 (-3.4)	9.7 (-3.9)
nC→OO	17.4 (-4.5)	9.3 (-4.8)

high-quality and precise semantic occupancy annotations for advancing 3D scene understanding. Additionally, the same model performs slightly better on nuCraft than on OpenOccupancy, indicating the consistency and reduced noise in our nuCraft dataset.

5.4 Cross-evaluation of VQ-Occ on OpenOccupancy and nuCraft

To assess the generalization ability of VQ-Occ, we perform cross-evaluation experiments on the OpenOccupancy and nuCraft datasets. When trained on OpenOccupancy and tested on nuCraft, the camera-based version of VQ-Occ experiences a performance drop of 3.9% in mIoU. Conversely, when trained on nuCraft and tested on OpenOccupancy, the performance drop is 4.8%. Table 7 presents the detailed cross-validation results. The performance drops observed in the cross-validation experiments can be attributed to the significant differences in the quality of the ground truth (GT) used for evaluation between OpenOccupancy and nuCraft, as illustrated in Fig. 1 OpenOccupancy GT suffers from several limitations, such as missing objects, noisy road boundaries, and thick ground planes. These inaccuracies and inconsistencies in the GT can negatively impact the evaluation of VQ-Occ's performance, regardless of which dataset it was trained on. When VQ-Occ is trained on OpenOccupancy and evaluated on nuCraft GT, which has cleaner and more precise annotations, the model's performance appears to degrade. This is because the model's predictions are being compared against a higher-quality GT, exposing the limitations of training on a dataset with noisy and inconsistent annotations. Conversely, when VQ-Occ is trained on nuCraft and evaluated on OpenOccupancy GT, the model's performance drops even further. This more significant drop can be attributed to the fact that the model, which was trained on cleaner and more accurate GT, is now being evaluated against a GT with more noise and inconsistencies. As a result, the model's predictions, which are likely to be more precise, are being penalized by the noisy and inconsistent GT used for evaluation.

5.5 Results on nuCraft@0.1

Predicting 3D semantic occupancy at a high resolution of 0.1m poses significant challenges due to the increased computational complexity and memory consumption. Despite these challenges, VQ-Occ demonstrates its potential by

Table 8: VQ-VAE reconstruction performance on the OpenOccupancy and nuCraft datasets at 0.2m resolution.

Dataset	IoU	mIoU	barrier	bicycle	bus	car	constr. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	vegetation
OpenOccupancy	68.6	49.2	52.3	45.1	48.6	50.8	47.2	44.5	46.9	48.0	50.1	53.7	59.4	55.8
nuCraft	73.1	50.7	54.0	46.6	50.1	52.4	48.6	45.8	48.3	49.5	51.7	55.3	61.2	57.5

directly generating predictions at 0.1m resolution without the need for post-processing or upsampling steps. Table 6 presents the preliminary results of VQ-Occ on nuCraft@0.1, along with a comparison of the estimated GPU memory consumption with C-CONet, a representative baseline method. Table 7 shows the cross-validation experiments on the OpenOccupancy and nuCraft datasets, highlighting the importance of using high-quality and consistent ground truth for accurate benchmarking.

These preliminary results demonstrate the effectiveness of VQ-Occ in tackling the challenges associated with high-resolution 3D semantic occupancy prediction. By directly generating predictions at 0.1m resolution while maintaining computational efficiency, VQ-Occ paves the way for more detailed and precise understanding of 3D scenes.

5.6 VQ-VAE Reconstruction

Here we assess the reconstruction quality of our VQ-VAE at 0.2m resolution in Table 8. Although the reconstruction performance may not seem exceptionally high, it is sufficient to serve as a robust foundation for providing codebooks for occupancy perception as VQ-Occ outperforms previous methods. This indicates the effectiveness of our approach in alleviating the learning difficulty compared to the direct learning of a large occupancy field. The current performance of our VQ-VAE also highlights the potential for further improvement in the codebook quality.

6 Conclusion

We introduce nuCraft, a high-resolution and precise 3D semantic occupancy dataset that provides high-resolution and precise annotations, enabling more detailed and accurate modeling of complex urban environments and paving the way for unifying 3D scene understanding. We also propose a novel 3D occupancy prediction framework VQ-Occ, which efficiently encodes high-resolution occupancy using VQ-VAE into a discrete latent space and learn 3D occupancy through feature simulation. Extensive experiments on OpenOccupancy and nuCraft datasets demonstrated the superiority of VQ-Occ over existing methods, setting new state-of-the-art performance for semantic occupancy prediction. We hope nuCraft and VQ-Occ will serve as a new challenging benchmark and baseline for future research on unifying 3D scene understanding, inspiring the development of advanced and efficient methods for high-resolution 3D semantic occupancy prediction in autonomous driving and beyond.

Acknowledgements

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by Smart Traffic Fund PSRI/76/2311/PR, by RGC General Research Fund Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019) 4
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 2, 4
3. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022) 12
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 5
5. Cheng, R., Agia, C., Ren, Y., Li, X., Bingbing, L.: S3cnet: A sparse semantic scene completion network for lidar point clouds. In: Conference on Robot Learning. pp. 2148–2161. PMLR (2021) 4
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 4
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 5
8. Fei, B., Yang, W., Chen, W.M., Ma, L.: Vq-dctr: Vector-quantized autoencoder with dual-channel transformer points splitting for 3d point cloud completion. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4769–4778 (2022) 5
9. Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J.: Structured prediction of unobserved voxels from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5431–5440 (2016) 4
10. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. IEEE Robotics and Automation Letters **7**(2), 3795–3802 (2022) 4
11. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: Scenenn: A scene meshes dataset with annotations. In: 2016 fourth international conference on 3D vision (3DV). pp. 92–101. Ieee (2016) 4
12. Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., Williams, F.: Neural kernel surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4369–4379 (2023) 6

13. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022) 10, 11
14. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021) 10, 11
15. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. arXiv preprint arXiv:2311.12754 (2023) 4
16. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9223–9232 (2023) 1, 12
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 2
18. Lee, J., Im, W., Lee, S., Yoon, S.E.: Diffusion probabilistic models for scene-scale 3d categorical data. arXiv preprint arXiv:2301.00527 (2023) 5
19. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9087–9098 (2023) 4
20. Li, Y., Dou, Y., Chen, X., Ni, B., Sun, Y., Liu, Y., Wang, F.: Generalized deep 3d shape prior via part-discretized diffusion process. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16784–16794 (2023) 5
21. Mersch, B., Guadagnino, T., Chen, X., Vizzo, I., Behley, J., Stachniss, C.: Building volumetric beliefs for dynamic environments exploiting map-based moving object segmentation. IEEE Robotics and Automation Letters (2023) 5
22. Pan, M., Liu, J., Zhang, R., Huang, P., Li, X., Liu, L., Zhang, S.: Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. arXiv preprint arXiv:2309.09502 (2023) 4
23. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019) 5
24. Roldao, L., de Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 2020 International Conference on 3D Vision (3DV). pp. 111–119. IEEE (2020) 12
25. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 746–760. Springer (2012) 4
26. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017) 4
27. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems **36** (2024) 1, 2, 3, 4, 7
28. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., Li, H.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8406–8415 (October 2023) 1, 2, 3, 4
29. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 3, 4, 9

30. Vizzo, I., Guadagnino, T., Mersch, B., Wiesmann, L., Behley, J., Stachniss, C.: KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. IEEE Robotics and Automation Letters (RA-L) **8**(2), 1029–1036 (2023). https://doi.org/10.1109/LRA.2023.3236571 6
31. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. arXiv preprint arXiv:2303.03991 (2023) 3, 4, 8, 12
32. Xia, Z., Liu, Y., Li, X., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y.: Scpnet: Semantic scene completion on point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17642–17651 (2023) 4
33. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE international conference on computer vision. pp. 1625–1632 (2013) 4
34. Xiong, Y., Ma, W.C., Wang, J., Urtasun, R.: Learning compact representations for lidar completion and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1074–1083 (2023) 5, 9, 11
35. Yan, J., Liu, Y., Sun, J., Jia, F., Li, S., Wang, T., Zhang, X.: Cross modal transformer: Towards fast and robust 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18268–18278 (2023) 5
36. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3101–3109 (2021) 12
37. Yu, Z., Shu, C., Deng, J., Lu, K., Liu, Z., Yu, J., Yang, D., Li, H., Chen, Y.: Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arXiv preprint arXiv:2311.12058 (2023) 4
38. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. arXiv preprint arXiv:2311.16038 (2023) 5
39. Zhong, X., Pan, Y., Behley, J., Stachniss, C.: Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8371–8377. IEEE (2023) 6
40. Zhu, B., Wang, Z., Shi, S., Xu, H., Hong, L., Li, H.: Conquer: Query contrast voxel-detr for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9296–9305 (2023) 7
41. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. arXiv preprint arXiv:2308.16896 (2023) 3, 4, 8, 12