COTR: Compact Occupancy TRansformer for Vision-based 3D Occupancy Prediction

Qihang Ma[1,*], Xin Tan[1,2,*], Yanyun Qu[3], Lizhuang Ma[1], Zhizhong Zhang[1,†], Yuan Xie[1,2]
[1] East China Normal University, Shanghai, China
[2] Chongqing Institute of East China Normal University, Chongqing, China
[3] Xiamen University, Fujian, China
qhma@stu.ecnu.edu.cn, {xtan,zzzhang,yxie}@cs.ecnu.edu.cn,
ma-lz@cs.sjtu.edu.cn, yyqu@xmu.edu.cn

Abstract

The autonomous driving community has shown significant interest in 3D occupancy prediction, driven by its exceptional geometric perception and general object recognition capabilities. To achieve this, current works try to construct a Tri-Perspective View (TPV) or Occupancy (OCC) representation extending from the Bird-Eye-View perception. However, compressed views like TPV representation lose 3D geometry information while raw and sparse OCC representation requires heavy but redundant computational costs. To address the above limitations, we propose Compact Occupancy TRansformer (COTR), with a geometry-aware occupancy encoder and a semantic-aware group decoder to reconstruct a compact 3D OCC representation. The occupancy encoder first generates a compact geometrical OCC feature through efficient explicit-implicit view transformation. Then, the occupancy decoder further enhances the semantic discriminability of the compact OCC representation by a coarse-to-fine semantic grouping strategy. Empirical experiments show that there are evident performance gains across multiple baselines, e.g.*, COTR outperforms baselines with a relative improvement of 8%-15%, demonstrating the superiority of our method. The code is available at https://github.com/NotACracker/COTR.*

1. Introduction

Vision-based 3D Occupancy Prediction aims to estimate the occupancy state of 3D voxels surrounding the ego-vehicle which provides a comprehensive 3D scene understanding [10, 32, 33, 37, 41]. By dividing the whole space into voxels and predicting its occupancy and semantic information, the 3D occupancy network endows a universal object representation ability, where out-of-vocabulary objects and abnormity can be easily represented as *[occupied; unknown]*.

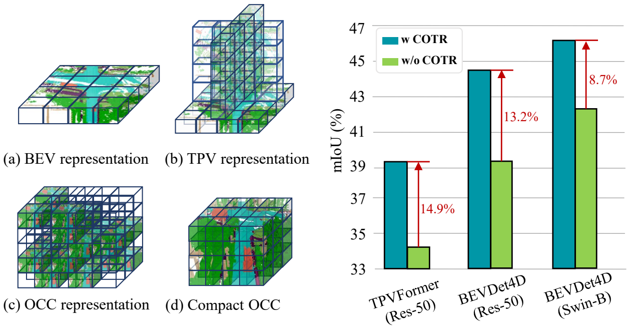

(a) BEV representation (b) TPV representation
(c) OCC representation (d) Compact OCC

Figure 1. **Left:** Different representation for 3D perception. **Right:** The 3D Occupancy prediction results of different baselines with COTR on nuScenes [32]. COTR outperforms baselines with a relative improvement of 8%-15%, demonstrating the superiority of our method.

3D vision perception is now transitioning from Bird-Eye-View (BEV) perception [9, 18, 21, 24, 26, 35] to Occupancy (OCC) perception [10, 25, 34, 37, 41]. The BEV perception excels in 3D object detection tasks due to their unified representation abilities for multi-camera inputs, where the obstruction problem is extremely alleviated in the BEV plane. However, its deficiency in collapsing the height dimensions poses a challenge in preserving the requisite geometric information for a holistic understanding of the 3D scene. To alleviate this issue, [10] proposes a Tri-Perspective View (TPV) representation to delineate 3D scenes. Unfortunately, this introduces a new problem in that the compression along the horizontal dimension leads to significant object overlap.

From previous empirical studies, it appears that the compression on a specific dimension of 3D representations would lose substantial 3D geometry information. Thus the idea of 3D OCC representation is quite intriguing. It is

* Equal contribution, † Corresponding author.

generated by dividing the 3D space into uniform grids and therefore mapping the 3D physical world into a 3D OCC representation. Obviously, this representation costs huge computational resources than previous BEV or TPV representations. Moreover, due to the sparsity, the information density of such uncompressed representation is low, with numerous regions corresponding to free space in the physical world, resulting in significant redundancy.

Another issue is that the current 3D OCC representation lacks semantic discriminability which impedes the network's ability to successfully recognize rare objects. This primarily stems from the problem of class imbalance within the dataset which is common in the field of autonomous driving. To substantiate this assertion, we conducted a straightforward proxy experiment. In particular, for the network's prediction, we maintained the occupancy prediction unchanged and substituted the semantic prediction for non-empty regions with corresponding ground-truth semantics. The experimental results indicate an improvement of about 95%, especially for the rare classes.

In this paper, we propose **C**ompact **O**ccupancy **TR**ansformer, termed as **COTR**, which aims to construct a compact 3D OCC representation. Our objective is to preserve rich geometric information and minimize computational costs while concurrently enhancing semantic discriminability.

In this framework, we propose to construct a compact geometry-aware 3D occupancy representation through efficient explicit-implicit view transformation. Specifically, after generating a sparse yet high-resolution 3D OCC feature by Explicit View Transformation (EVT), we downsampled it to a compact OCC representation, a size merely 1/16 of the original, without any performance drop. Taking the compact OCC feature as input, Implicit View Transformation (IVT) further enriches it through spatial cross-attention and self-attention. Then, the updated OCC feature is upsampled to the original resolution for the downstream module. To recover the geometric details lost during downsampling, we configure the downsampling and upsampling processes into a U-Net architecture. Through such an approach, we substantially mitigate the sparsity of the OCC feature while simultaneously retaining geometric information and reducing unnecessary computational overhead and training time introduced by IVT.

Secondly, we introduce a coarse-to-fine semantic-aware group decoder. We first divide the ground-truth labels into several groups based on semantic granularity and sample count. Then, for each semantic group, we generate corresponding mask queries and train the network based on group-wise one-to-many assignment. The grouping strategy results in balanced supervision signals, significantly enhancing the capability to recognize different classes, leading to a compact geometry- and semantic-aware OCC representation.

Our contributions can be summarized as follows:

- We propose a geometry-aware occupancy encoder to construct a compact occupancy representation through efficient explicit-implicit view transformation. We can handle the sparsity of the occupancy feature while preserving the geometry information and reducing computation costs.
- We proposed a novel semantic-aware group decoder that significantly enhances the semantic discriminability of the compact occupancy feature. This group strategy balances the supervision signals and alleviates the suppression from common objects to rare objects.
- Our method has been embedded into several prevailing backbones, and experiments on the Occ3D-nuScenes dataset show that our approach achieves state-of-the-art performance. What's more, our method outperforms the backbones with a relative improvement of 8%-15%, as illustrated in Fig. 1.

2. Related Work

2.1. Vision-based BEV Perception

Over the recent years, vision-based Bird-Eye-View (BEV) perception has undergone significant development [12], emerging as a crucial component in the autonomous driving community due to its cost-effectiveness, stability, and versatility. By transforming 2D image features to a unified and comprehensive 3D BEV representation through view transformation, various tasks, including 3D object detection and map segmentation, have been consolidated within a unified framework. View transformation can be broadly categorized into two types: one relies on explicit depth estimation to form a pseudo point cloud and construct the 3D space [8, 16, 20, 26, 27], while the other pre-defines the BEV space and implicitly models the depth information through spatial cross-attention [11, 18, 35, 36, 40], mapping image features into the corresponding 3D positions. Although BEV perception excels in 3D object detection, it still encounters challenges in handling corner cases in driving scenarios, including irregular obstacles and out-of-vocabulary objects. To alleviate the aforementioned challenges, the 3D occupancy prediction task was proposed.

2.2. 3D Occupancy Prediction

The 3D occupancy prediction task has garnered significant attention due to its enhanced geometry information and superior capabilities in generalized object recognition compared to 3D object detection. TPVFormer [10] adopts the concept of BEV perception, dividing 3D space into three perspective views and utilizing sparse point cloud supervision for 3D occupancy prediction. SurroundOcc [37] generates geometric information by expanding the heightdimensional of the BEV feature into occupancy features and conducting spatial cross-attention directly on them. Additionally, they introduce a new pipeline for constructing occupancy ground truth. OccNet [33] bridges the end-to-end framework from perception to planning by constructing a general occupancy embedding. FBOcc [19] proposed a forward-backward view transformation module based on the BEV feature to address the limitations of different view transformations. While the aforementioned methods have made initial strides in the occupancy prediction task, a majority of them have largely adhered to the BEV perception framework and straightforwardly transformed BEV features to OCC features for the final prediction. They do not consider the sparsity and lack of semantic discriminability in the raw OCC representation.

2.3. Semantic Scene Completion

The definition of 3D occupancy prediction shares the most resemblance with Semantic Scene Completion (SSC) [5, 13, 28, 30, 39]. MonoScene [3] first proposed a framework that inferred dense geometry and semantics from a single monocular 2D RGB image. Voxformer [17] draws on the idea of BEV perception and employs depth estimation to construct a two-stage framework, which mitigates the overhead linked to attention computation. Occformer [41] proposed a dual-path transformer and adopted the concept of mask classification for occupancy prediction. However, the performance of Voxformer relies on the robustness of the depth estimation, whereas Occformer's utilization of various transformers significantly increases the number of parameters. In this paper, we introduce an efficient framework to boost the performance of occupancy prediction while maintaining a low computational cost.

3. Methodology

3.1. Preliminary

Given a sequence of multi-view image inputs, the goal of vision-centric 3D occupancy prediction is to estimate the state of 3D voxels surrounding the ego-vehicle. Specifically, the input of the task is a T-frame consequent sequence of images $\{I_{i,t} \in \mathbb{R}^{H_i \times W_i \times 3}\}$ from N_c surround-view cameras, where $i \in \{1, \dots, N_c\}$ and $t \in \{1, \dots, T\}$. Besides, the camera intrinsic parameters $\{K_i\}$ and extrinsic parameters $\{[R_i|t_i]\}$ are also known for coordinate system conversions and ego-motion.

3D occupancy prediction aims to infer the states of each voxel, including *occupancy* (*[occupied]* or *[empty]*) and *semantics* (*[category]* or *[unknown]*) information. For example, a voxel on a car is annotated as *[occupied; car]*, and a voxel in the free space is annotated as *[empty; None]*. One primary advantage of the 3D occupancy prediction is to provide a universal object representation, where out-of-vocabulary objects and abnormity can be easily represented as *[occupied; unknown]*.

3.2. Overall Architecture

An overview of the Compact Occupancy TRansforemr (COTR) is presented in Fig. 2. The COTR mainly consists of three key modules: an image featurizer to extract image features and depth distributions, a geometry-aware occupancy encoder (Sec. 3.3) to generate a compact occupancy representation through efficient explicit-implicit view transformation, and a semantic-aware group decoder (Sec. 3.4) to further enhance the semantic discriminability and geometry details of the compact OCC feature.

Image Featurizer. The image featurizer aims to extract image features and depth distributions for multi-camera inputs, which provides the foundation of the geometry-aware occupancy encoder. Given a set of RGB images from multiple cameras, we first use a pretrained image backbone network (*e.g.*, ResNet-50 [7]) to extract image features $F = \{F_i \in \mathbb{R}^{C_F \times H \times W}\}_{i=1}^{N_c}$, where F_i is the view features of i-th camera view and N_c is the total number of cameras. Next, the depth distributions $D = \{D_i \in \mathbb{R}^{D_{bin} \times H \times W}\}_{i=1}^{N_c}$ can be obtained by feeding these image features F into depth net.

3.3. Geometry-aware Occupancy Encoder

A key insight behind the occupancy task is that it could capture the fine-grained details of critical obstacles in the scene, such as the geometric structure of an object. To do so, we decided to use both explicit and implicit view transformation to generate a compact geometry-aware occupancy representation. In this section, we will first briefly review the explicit-implicit view transformation, and then intricately delineate how to construct the compact occupancy representation through efficient explicit-implicit fusion.

Explicit-implicit View Transformation. Explicit and Implicit View Transformation is a crucial step in BEV perception [9, 18, 26, 35] to transform 2D image features to BEV representation. In order to construct a 3D representation that can preserve more 3D geometry information, we extend the Explicit-Implicit VT for OCC representation construction. Specifically, for EVT, the image features F and depth distributions D are computed by outer product $F \otimes D$ to obtain a pseudo point cloud feature $P \in \mathbb{R}^{N_c D_{bin} HW \times C}$. Then, instead of creating a BEV feature $B_\mathrm{E} \in \mathbb{R}^{C \times X \times Y}$, we directly generate a 3D OCC feature $O_\mathrm{E} \in \mathbb{R}^{C \times X \times Y \times Z}$ through voxel-pooling, where (X, Y, Z) denotes the resolution of the 3D volume. For IVT, we predefine a group of grid-shaped learnable parameters $Q \in \mathbb{R}^{C \times X \times Y \times Z}$ as the queries of OCC, where each query is responsible for the corresponding grid cell in the 3D Occupancy space. Then the set of OCC queries will be updated through spatial cross-attention and self-attention to

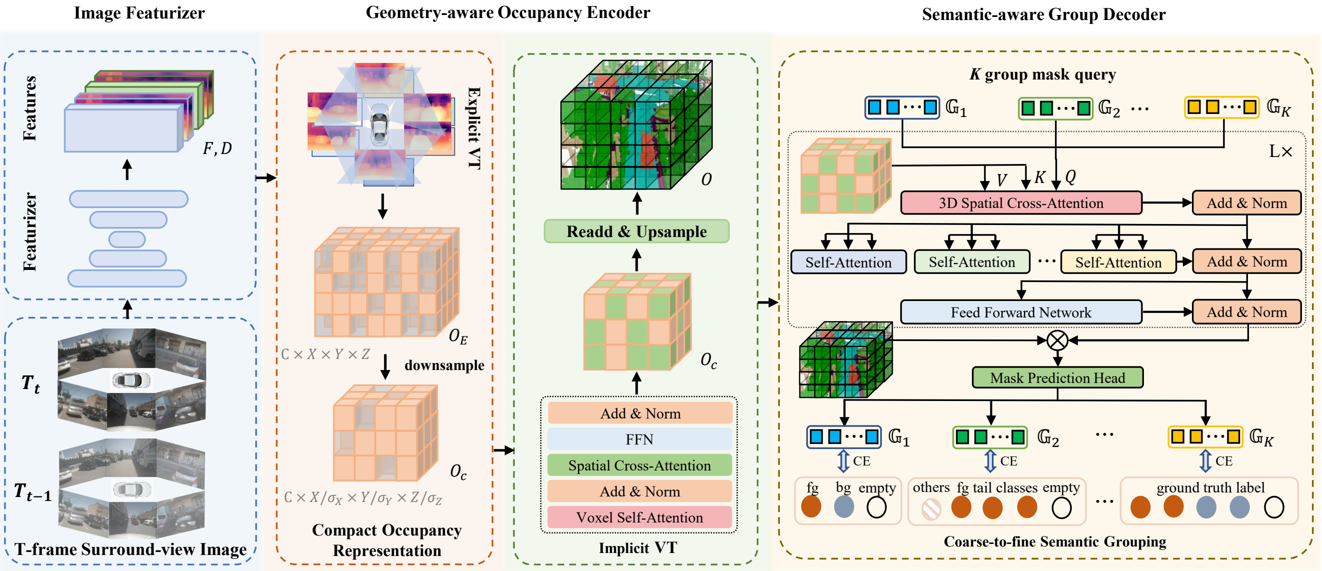


Figure 2. The overall architecture of COTR. T-frame surround-view images are first fed into the image featurizers to get the image features and depth distributions. Taking the image features and depth estimation as input, the geometry-aware occupancy encoder constructs a compact occupancy representation through efficient explicit-implicit view transformation. The semantic-aware group decoder utilizes a coarse-to-fine semantic grouping strategy cooperating with the Transformer-based mask classification to strongly strengthen the semantic discriminability of the compact occupancy representation.

interact with the image features.

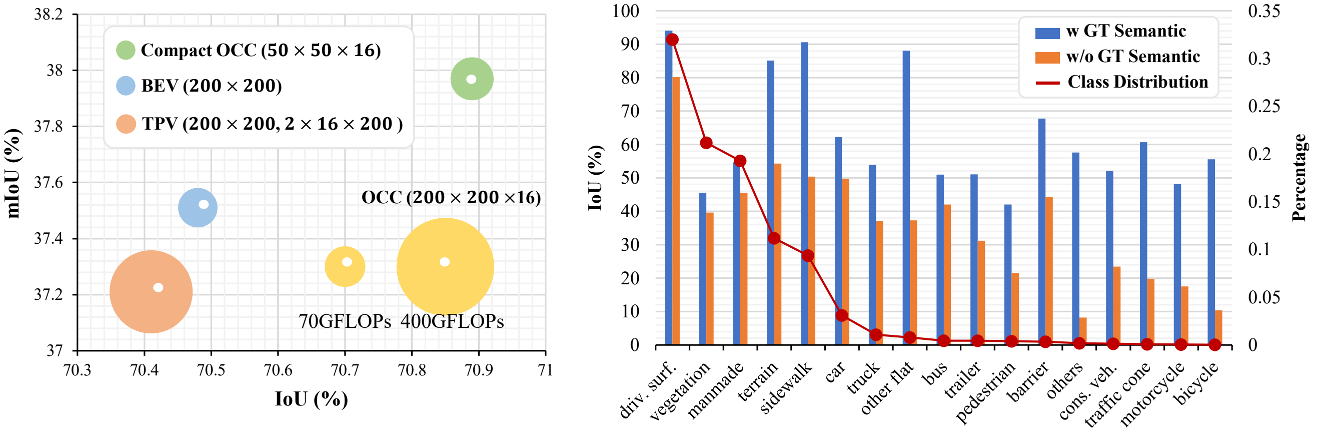
Compact Occupancy Representation. With the adoption of EVT, we have already obtained a geometry-aware 3D OCC feature. A straightforward fusion approach is to directly incorporate this feature as the input to IVT. However, computing SCA with high-resolution 3D OCC features (*e.g.*, $200 \times 200 \times 16$) incurs significant computational overhead. Additionally, due to the sparsity of 3D space, the computation in the majority of free space is also ineffective. Therefore, we downsampled the high-resolution but sparse OCC feature O_{E} to a compact OCC representation $O_c \in \mathbb{R}^{C\times\frac{X}{\sigma_x}\times\frac{Y}{\sigma_y}\times\frac{Z}{\sigma_z}}$, where $\sigma_x, \sigma_y, \sigma_z$ denote the downsample ratios. Taking each 3D voxel of the compact OCC O_c as queries, IVT completes the sparse regions and further enriches the geometric details therein. Compared to a standard encoder which learns a set of queries from scratch, this operation significantly saves extra training time and reduces computational overhead. Then the compact O_c is upsampled to original resolution $O \in \mathbb{R}^{C_Q\times X\times Y\times Z}$ for final occupancy prediction. Since the downsampling operation inevitably introduces information loss, especially for small objects, we constructed the downsampling and upsampling processes into a Unet [29] architecture to mitigate this problem. In practice, we construct a compact OCC representation with a size merely 1/16 of the original while achieving a better performance.

Discussion. There are three overarching advantages to employing the compact occupancy representation. Firstly, the 3D feature representation enjoys a natural geometric superiority over 2D BEV or TPV [10]. As illustrated in Fig. 3 (a), the compact OCC representation achieves the best IoU score. Secondly, the compact OCC representation effectively alleviates the sparsity inherent in high-resolution OCC features. For outdoor autonomous driving datasets such as Nuscenes [1], SemanticKITTI [30], and Waymo [31], the proportion of free space stands at 78%, 93%, and 92%, respectively. The compact OCC representation denotes a compressed spatial domain, enriched features, and expanded receptive fields. Thirdly, the computational overhead is significantly diminished. As depicted in Fig. 3 (a), the computational cost of a raw high-resolution OCC representation is approximately 500% higher than that of a compact OCC representation.

3.4. Semantic-aware Group Decoder

In this section, we present our semantic-aware group decoder, which further enhances the geometric occupation of the compact OCC feature while greatly improving semantic discriminability. We will commence with a proxy experiment designed to substantiate our assertion that the occupancy feature lacks semantic discriminability, which significantly impedes the recognition of rare objects. Subsequently, we will delve into the details of our coarse-to-fine semantic grouping strategy.

Proxy Experiment. To demonstrate that the occupancy feature lacks semantic discriminability, we replace the semantic prediction in our occupancy prediction with the ground-truth label in the corresponding position. As shown in Fig. 3 (b), the IoU scores are greatly improved, especially for the tail class. This drives us to look for a new approach



(a) Comparison of different representation.

(b) Per-class mIoU and distribution.

Figure 3. Proxy experiments. (a) depicts the comparison of different occupancy representations. The compact 3D OCC representation achieves a balance between performance and computational cost. (b) reports the per-class mIoU and distribution with and without using ground-truth semantic labels.

to greatly enhance the semantic discrimination of the occupancy feature.

Transformer Decoder. Inspired by MaskFormer [6], We convert the occupancy prediction to the form of mask classification. This form of prediction splits the occupancy prediction into two sub-problems, which is convenient for us to address the semantic obfuscation problem. To do so, we replace the image feature in the Transformer decoder with the compact occupancy feature O_c from our geometry-aware occupancy encoder. In addition, we replace the original encoder-decoder global self-attention layer with 3D SCA to further reduce the computational cost. The 3D Spatial Cross Attention (3D-SCA) can be formulated as:

$$\text{3D-SCA}(Q_m, O_c) = \sum_{i=1}^{\mathcal{N}_{\text{ref}}} f(Q_m, \mathbf{p}_i, O_c), \qquad (1)$$

where $f(\cdot)$ is the deformable attention function, $Q_m \in \mathbb{R}^{N_q \times C_m}$ is N_q learnable mask queries, $\mathbf{p}$ denotes the sample location in the 3D occupancy space and we sample $\mathcal{N}_{\text{ref}}$ 3D occcupancy features for each mask query Q_m. The updated mask queries are then fed into self-attention to interact with other mask queries. At the end of each iteration, each mask query $q_i \in Q_m$ will be projected to predict its class probability $\{p_i \in \delta^{K+1}\}_{i=1}^{N_q}$ and the mask embedding $\mathcal{E}_m$. Then, the latter is further converted to a binary occupancy mask $m_i \in [0,1]^{X \times Y \times Z}$ through a dot product with the occupancy feature O and a sigmoid function:

$$m_i[x,y,z] = \text{sigmoid}(\mathcal{E}_m[:,i]^{\text{T}} \cdot O[:,x,y,z]). \qquad (2)$$

Coarse-to-Fine Semantic Grouping. Because of the imbalanced data distribution, the classifier that predicts the class probability would make the classification scores for low-shot categories much smaller than those of many-shot categories, resulting in semantic misclassification. In order to enhance the supervision signal of the rare classes, we first adopt a group-wise one-to-many assignment according to [4], which aims to enable each mask query to obtain multiple positive matching pairs. However, experiments show that such a simple grouping strategy is ineffective and can not bring performance improvement. Inspired by [2, 14, 38], based on the group-wise one-to-many assignment, we further introduce a Coarse-to-Fine Semantic Grouping strategy. This involves partitioning mask queries into K groups, with each group supervised by dividing semantic categories into K ground truth (gt) label groups based on semantic granularity and sample count, aimed at balancing the supervision signal in each group.

Concretely, we first divided the categories into $\mathbb{G}_1 = \{$*"foreground", "background", "empty"*$\}$. Then for the foreground or background category, we again grouped the categories based on the number of training samples. We assign category i into group $\mathbb{G}_n$ if:

$$N_n^l < \mathcal{N}(i) \leq N_n^h, \quad n \in [2, K-1], \qquad (3)$$

where $\mathcal{N}(i)$ is the number of training samples for category i, and N_n^l and N_n^h are hyper-parameters that determine minimal and maximal sample numbers for group n.

In order to keep only one group during inference, we manually set the $\mathbb{G}_K$ to the original ground truth label and $\mathbb{G}_1$ to $\{$*"foreground", "background", "empty"*$\}$, thus there is no extra cost compared with single-group models. Besides, during training, in order for each class to be supervised in each group, we categorize the classes not assigned to the current group as *"others"*. For example, if the foreground classes *motorcycle* and *bicycle* are assigned in the same group, the ground truth label for this group is $\{$*"others", "motorcycle", "bicycle", "empty"* $\}$. For label

Method	Venue	Image Backbone	Image Size	Epoch	Visible Mask	IoU (%)	mIoU (%)
MonoScene [3]	CVPR'22	ResNet-101	928 × 600	24	✘	-	6.1
OccFormer* [41]	ICCV'23	ResNet-50	256 × 704	24	✘	30.1	20.4
BEVFormer [18]	ECCV22	ResNet-101	928 × 600	24	✘	-	26.9
CTF-Occ [32]	arXiv'23	ResNet-101	928 × 600	24	✘	-	28.5
VoxFormer [17]	CVPR'23	ResNet-101	900 × 1600	24	✔	-	40.7
SurroundOcc [37]	ICCV'23	InternImage-B	900 × 1600	24	✔	-	40.7
FBOcc† [19]	ICCV'23	ResNet-50	256 × 704	20	✔	-	42.1
TPVFormer* [10]	CVPR'23	ResNet-50	900 × 1600	24	✔	66.8	34.2
TPVFormer + COTR	-	ResNet-50	256 × 704	24	✔	**70.6**	**39.3**
SurroundOcc* [37]	ICCV'23	ResNet-101	900 × 1600	24	✔	65.5	34.6
SurroundOcc + COTR	-	ResNet-50	256 × 704	24	✔	**71.0**	**39.3**
OccFormer* [41]	ICCV'23	ResNet-50	256 × 704	24	✔	70.1	37.4
OccFormer + COTR	-	ResNet-50	256 × 704	24	✔	**71.7**	**41.2**
BEVDet4D‡ [8]	arXiv'22	ResNet-50	384 × 704	24	✔	73.8	39.3
BEVDet4D + COTR	-	ResNet-50	256 × 704	24	✔	**75.0**	**44.5**
BEVDet4D‡ [8]	arXiv'22	SwinTransformer-B	512 × 1408	36	✔	72.3	42.5
BEVDet4D + COTR	-	SwinTransformer-B	512 × 1408	24	✔	**74.9**	**46.2**

Table 1. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset.** † with test-time augmentation. ‡ means the performance is reported by its official code. * means the performance is achieved by our implementation using its official code. Visible Mask means whether models are trained with visible masks.

groups like $\mathbb{G}_1$, category categorization is coarse, while in the subsequent groups, it becomes more fine-grained.

4. Experiments

4.1. Experimental Setup

Dataset. Occ3D-nuScenes [32] is a large-scale autonomous driving dataset, which contains 700 training scenes and 150 validation scenes. Each frame contains a 32-beam LiDAR point cloud and six RGB images captured by six cameras from different views of LiDAR with dense voxel-wise semantic occupancy annotations. The occupancy scope is defined as -40m to 40m for the X and Y axis, and -1m to 5.4m for the Z axis in the ego coordinate. The voxel size is 0.4m × 0.4m × 0.4m for the occupancy label. The semantic labels contain 17 categories, including 16 known object classes with an additional "empty" class.

Implementation Details. By adhering to common practices [9, 15, 19], we default to using ResNet-50 [7] as the image backbone, and the image size is resized to (256×704) for Occ3D-nuScenes. For explicit view transformation, we adopt BEVStereo [15] which depth estimation is supervised from sparse LiDAR. The resolution of the occupancy feature from explicit view transformation is 200×200×16 with a feature dimension C of 32, and the downsample ratios are $\sigma_x = \sigma_y = 4, \sigma_z = 1$ with an embedding dimension C_Q of 256. We use 8 attention heads for both self- and cross-attention and set $\mathcal{N}_{ref} = 4$ for both 2D and 3D SCA. We simply generate $K = 6$ group gt labels and mask queries for the Group Occupancy Decoder, which We simply divided both the foreground and background classes into two separate groups, resulting in a total of 4 groups. This division was based on the median number of training samples.

Component			Metric			
GOE	TD	CFSG	IoU (%)		mIoU (%)	
			70.36	-	36.01	-
✔			70.89	+0.53	37.97	+1.98
	✔		69.76	-0.60	38.43	+2.42
✔	✔		71.74	+1.38	40.22	+4.21
✔	✔	✔	**72.08**	**+1.72**	**41.39**	**+5.38**

Table 2. **Ablation study on the each component.** GOE denotes the Geometry-aware Occupancy Encoder, TD denotes the Transformer Decoder and CFSG means Course-to-Fine Semantic Grouping. All models are trained without long-term temporal information.

We also integrate our method into four main-stream occupancy models BEVDet4D [8], TPVFormer [10], SurroundOcc [37] and OccFormer [41] to demonstrate the effectiveness of our approach. Unless otherwise specified, all models are trained for 24 epochs using AdamW optimizer [23], in which gradient clip is exploited with learning rate 2e-4.

4.2. Comparing with SOTA methods

Occ3D-nuScenes. As shown in Table 1, we report the quantitative comparison of existing state-of-the-art methods for 3D occupancy prediction tasks on Occ3D-nuScenes. We integrate our method into TPVFormer, SurroundOcc, OccFormer and BEVDet4D, and our approach yields significant performance improvements in both geometric completion and semantic segmentation, surpassing the baseline by 3.8%, 5.5%, 1.6%, 1.2% in IoU and 5.1%, 4.7%, 3.8%, 5.2% in mIoU. Notably, our approach based on BEVDet4D

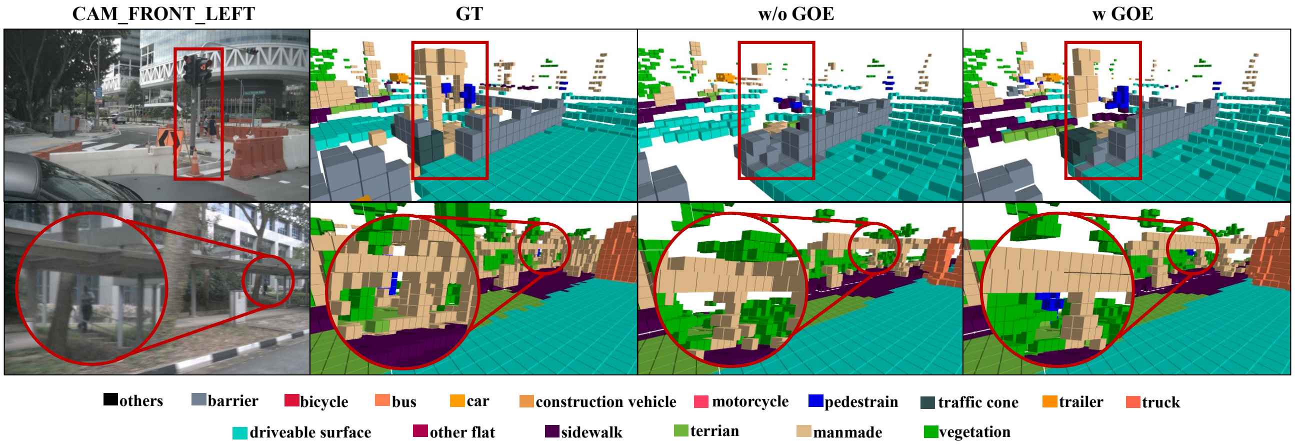

CAM_FRONT_LEFT
GT
w/o GOE
w GOE
■others ■barrier ■bicycle ■bus ■car ■construction vehicle ■motorcycle ■pedestrian ■traffic cone ■trailer ■truck
■driveable surface ■other flat ■sidewalk ■terrian ■manmade ■vegetation

Figure 4. Qualitative results comparison between baseline and our Geometry-aware Occupancy Encoder. The results demonstrate that the compact occupancy representation is able to capture more precise geometrical details for slimmer objects (*e.g.*, pedestrians and poles) and is robust to occlusions by combining the advantages of implicit and explicit view transformation.

Groups	LT	IoU (%)		mIoU (%)	
{*gt label*}	✘	71.74	-	40.22	-
10× {*gt label*}	✘	71.47	-0.27	40.17	-0.05
{*fg*, *bg*, *empty*}, {*gt label*}	✘	72.11	+0.37	40.61	+0.39
{*fg*, *bg*, *empty*}, . . ., {*gt label*}	✘	72.08	+0.34	41.39	+1.17
{*gt label*}	✔	73.68	-	42.70	-
{*fg*, *bg*, *empty*}, . . ., {*gt label*}	✔	75.01	+1.33	44.45	+1.75

Table 3. **Ablation study on the number of Semantic Groups.** 10× {*gt label*} means we replicated the original ground truth labels ten times to verify that the performance improvement was due to the increase in model parameters. LT means models are trained with long-term temporal information.

utilizes a smaller backbone (ResNet-50) and smaller image input size (256 × 704) to achieve mIoU scores of 44.5%, which outperforms both Voxformer [17] (ResNet-101, 900 × 1600) and SurroundOcc [37] (InternImage-B) by 3.8%. This demonstrates that our method mines more information with a small number of parameters by designing components specifically for 3D occupancy prediction tasks. Compared to the state-of-the-art FBOcc [19], BEVDet4D with COTR surpasses it by 2.3% even without test-time augmentation. In addition, we also scale up the image backbone to SwinTransformer-B [22] and Image Size to 512 × 1408. The experimental results demonstrate that our method consistently brings performance improvements, even with larger model sizes.

4.3. Ablation study

To delve into the effect of different modules, we conduct ablation experiments on Occ3d-nuScenes [32] based on BEVDet4D [8].

The Effectiveness of Each Component. The results are shown in Table 2, we can observe that all components make their own performance contributions. The baseline achieves 70.36% of IoU and 36.01% of mIoU without long-term temporal information. We first integrated the Geometry-aware Occupancy Encoder (GOE) into the baseline model, which brings 0.53% and 1.98% performance gain in IoU and mIoU. By converting the 3D occupancy task to mask classification with the help of a Transformer decoder (TD), the network's semantic segmentation capability has been significantly enhanced. By using both GOE and TD, the network can excel in both geometric completion and semantic segmentation, outperforming the baseline by 1.38% of IoU and 4.21% of mIou. Moreover, the Coarse-to-Fine Semantic Grouping further enhances rare object recognition and achieves 41.05% mIoU scores while retaining essentially the same geometry completion capability.

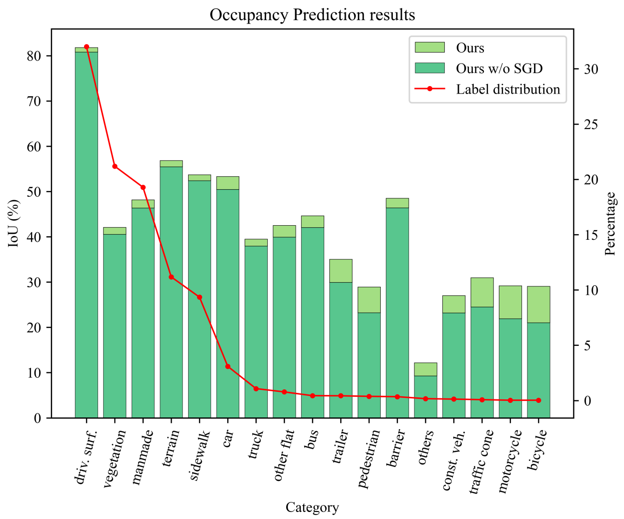


Figure 5. The occupancy prediction results with label distribution. It is clear to see that the Semantic-aware Group Decoder (SGD) gives a big performance boost to the rare class.

The Effectiveness of Compact Occupancy Representation. To further demonstrate the effect of a compact occupancy representation, we conducted an experiment where

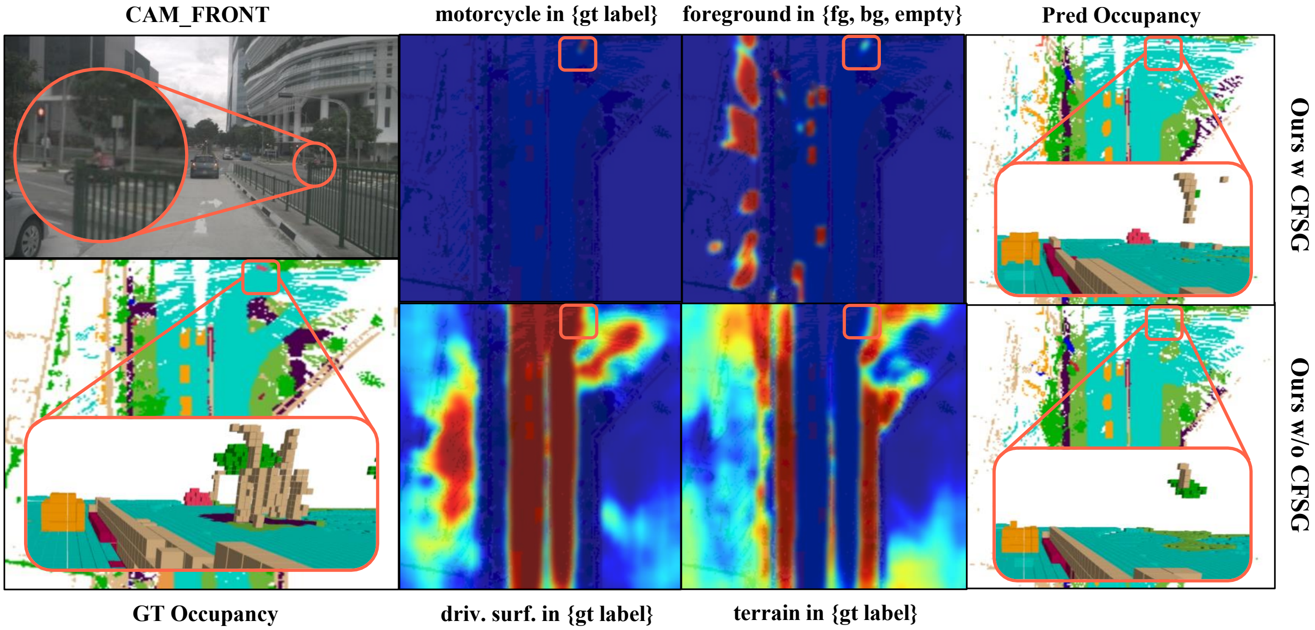


Figure 6. We visualize the heatmaps using different group masks. *motorcycle in* {*gt label*} means the mask query is supervised by ground truth label, and *foreground in* {*fg, bg, empty*} means the mask query is supervised by semantic group {*"foreground", "background", "empty"*}. The rare category *motorcycle* was successfully detected in various semantic groups, thereby enhancing the supervision signal.

we used different representations of the occupancy features. As shown in Fig. 3 (a), by introducing a U-net to bridge the explicit and implicit view transformation in a voxel representation, we have achieved a balance between performance and computational efficiency. Qualitative results illustrated in Fig. 4 showcase that the compact occupancy representation is able to bring improvements in geometric completion, especially for slender objects such as pedestrians and poles. Moreover, the compact occupancy representation exhibits robustness to occlusions.

The Effectiveness of Semantic-aware Group Decoder. In Fig. 5, we compare the results of adopting the Semantic-aware Group Decoder (SGD) according to the label distribution. It is clear to see that there exists a significant class imbalance phenomenon in the dataset, for example, where the 6 background categories account for 93.8% of the total labels. SGD significantly enhances the semantic discriminability of the compact occupancy representation through the transformer decoder and balancing supervision within each group using coarse-to-fine semantic grouping.

The Effectiveness of Coarse-to-Fine Semantic Grouping. To further demonstrate the impact of CFSG, we compare the performance impact of using different numbers of semantic groups. As shown in Table 3, replicating the original ground truth labels ten times like [4] to formulate a one-to-many assignment and increase the number of model parameters does not result in performance improvement. However, adding a simple {*"foreground", "background", "empty"*} group can give a significant performance boost. Moreover, we generated heatmaps to visualize query masks in various groups. As illustrated in Fig. 6, the rare class motorcycle is correctly detected in both the group of {*gt label*} or {*"foreground", "background", "empty"*}, which strengthens the semantic supervision of this class. In contrast, the network without using CFSG experienced suppression of its response values by the background class at that position.

5. Conclusion

In this paper, we have presented COTR, a Compact Occupancy TRansformer for vision-based 3D occupancy prediction. For a holistic understanding of the 3D scene, we construct a compact geometry- and semantic-aware 3D occupancy representation through efficient explicit-implicit view transformation and coarse-to-fine semantic grouping. We evaluate COTR with several prevailing baselines and achieve state-of-the-art performance on nuScenes. We hope COTR can motivate further research in vision-based 3D occupancy prediction and its applications in autonomous vehicle perception.

Acknowledgements This work is supported by the National Natural Science Foundation of China No.62302167, U23A20343, 62222602, 62206293, 62176224, 62106075, and 62006139, Shanghai Sailing Program (23YF1410500), Natural Science Foundation of Shanghai (23ZR1420400), Science and Technology Commission of Shanghai No.21511100700, Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137), CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (RAGR20230121) and CAAI-Huawei MindSpore Open Fund.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11621–11631, 2020. 4

[2] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 112–121, 2021. 5

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

[4] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6633–6642, 2023. 5, 8

[5] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4193–4202, 2020. 3

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems (NIPS)*, 34:17864–17875, 2021. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 3, 6

[8] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2, 6, 7

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

[10] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9223–9232, 2023. 1, 2, 4, 6

[11] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1042–1050, 2023. 2

[12] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022. 2

[13] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3351–3359, 2020. 3

[14] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10991–11000, 2020. 5

[15] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1486–1494, 2023. 6

[16] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1477–1485, 2023. 2

[17] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, 2023. 3, 6, 7

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision (ECCV)*, pages 1–18. Springer, 2022. 1, 2, 3, 6

[19] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6919–6928, 2023. 3, 6, 7

[20] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems (NIPS)*, 35:10421–10434, 2022. 2

[21] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision (ECCV)*, pages 531–548. Springer, 2022. 1

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 10012–10022, 2021. 7

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Chenyang Lu, Marinus Jacobus Gerardus Van De Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–

decoder networks. *IEEE Robotics and Automation Letters*, 4 (2):445–452, 2019. 1

[25] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 1

[26] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision (ECCV)*, pages 194–210. Springer, 2020. 1, 2, 3

[27] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8555–8564, 2021. 2

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

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 4

[30] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1754, 2017. 3, 4

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2446–2454, 2020. 4

[32] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 1, 6, 7

[33] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8406–8415, 2023. 1, 3

[34] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17850–17859, 2023. 1

[35] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 1, 2, 3

[36] Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Frustumformer: Adaptive instance-aware resampling for multi-view 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5096–5105, 2023. 2

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

[38] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European conference on computer vision (ECCV)*, pages 247–263. Springer, 2020. 5

[39] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3101–3109, 2021. 3

[40] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17830–17839, 2023. 2

[41] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9433–9443, 2023. 1, 3, 6