H2GFormer: Horizontal-to-Global Voxel Transformer for 3D Semantic Scene Completion

Yu Wang[1], **Chao Tong**[2,3]*

[1]Sino-French Engineer School, Beihang University
[2]School of Computer Science and Engineering, Beihang University
[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
{wangyu0089, tongchao}@buaa.edu.cn

Abstract

3D Semantic Scene Completion (SSC) has emerged as a novel task in vision-based holistic 3D scene understanding. Its objective is to densely predict the occupancy and category of each voxel in a 3D scene based on input from either LiDAR or images. Currently, many transformer-based semantic scene completion frameworks employ simple yet popular Cross-Attention and Self-Attention mechanisms to integrate and infer dense geometric and semantic information of voxels. However, they overlook the distinctions among voxels in the scene, especially in outdoor scenarios where the horizontal direction contains more variations. And voxels located at object boundaries and within the interior of objects exhibit varying levels of positional significance. To address this issue, we propose a transformer-based SSC framework called H2GFormer that incorporates a horizontal-to-global approach. This framework takes into full consideration the variations of voxels in the horizontal direction and the characteristics of voxels on object boundaries. We introduce a horizontal window-to-global attention (W2G) module that effectively fuses semantic information by first diffusing it horizontally from reliably visible voxels and then propagating the semantic understanding to global voxels, ensuring a more reliable fusion of semantic-aware features. Moreover, an Internal-External Position Awareness Loss (IoE-PALoss) is utilized during network training to emphasize the critical positions within the transition regions between objects. The experiments conducted on the SemanticKITTI dataset demonstrate that H2GFormer exhibits superior performance in both geometric and semantic completion tasks. Our code is available on https://github.com/Ryanwy1/H2GFormer.

Introduction

Holistic 3D understanding has emerged as a crucial challenge in computer vision, advancing notably in scenarios like autonomous driving, robot behavior planning, and virtual reality applications. However, occlusions and incomplete observations pose challenges in acquiring precise 3D information.

To address these challenges, SSCNet (Song et al. 2017) pioneered the introduction of 3D semantic scene completion and demonstrated scene completion and semantic labeling within this task are tightly intertwined. Subsequent methods (Li et al. 2019a; Garbade et al. 2019; Li et al. 2020b) continue this perspective and place a greater emphasis on indoor scenes. Other works (Roldao, de Charette, and Verroust-Blondet 2020; Cheng et al. 2021; Yan et al. 2021; Yang et al. 2021; Xia et al. 2023) leverage 3D geometric data, such as LiDAR point clouds, extending this task to outdoor scenes. Compared to the relatively higher cost of LiDAR sensors, visual cameras not only offer cost-effective solutions but also provide abundant visual cues of the scene. It has prompted recent research endeavors to shift towards vision-based solutions. MonoScene (Cao and de Charette 2022) was the pioneering approach that introduced semantic scene completion utilizing monocular images as input. Other works (Huang et al. 2023; Li et al. 2023) aimed to improve the model's ability to generate dense 3D scenes from sparse observations by refining the design of the network architecture.

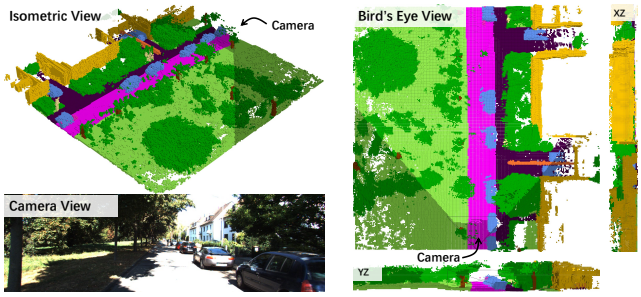

Figure 1: The isometric view, camera view, and bird's-eye view of SemanticKITTI (Cai et al. 2021). In outdoor scenes, voxel categories exhibit greater variation along the horizontal direction than along the vertical ground direction.

However, these Transformer-based visual approaches predominantly focus on holistic voxel feature modeling, thereby overlooking distinctions between voxels along the horizontal and vertical directions, as well as the disparities between voxels within objects and those outside. Hence, these methods encounter inherent challenges such as the inability to accurately and efficiently diffuse 2D image features into 3D voxels. In reality, the uniform treatment of all voxels is not tenable, especially in outdoor scenarios where the horizontal dimension exhibits more variations, as shown

*Corresponding author.

in Figure 1. In addition to the presence of more variations in object categories, there is also a considerable difference in the scales of objects in the horizontal direction. This requires the model to be capable of modeling longer-distance contextual semantics. And the outdoor images encompass a wider range of the real-world environment. This accentuates the imperative for dependable exploitation of semantic features at object edges.

Motivated by the above analysis, we propose a horizontal-to-global voxel reconstruction network (H2GFormer) that leverages 2D image features more reliably and efficiently, thereby achieving semantic scene completion. The key idea of H2GFormer involves a gradual focus on the reconstruction of voxel features at diverse directions and positions. Initially, we perform extensive contextual semantic modeling along the horizontal direction, subsequently expanding local information to encompass height-related regions. Consequently, H2GFormer achieves a more dependable and precise diffusion of 2D image features into 3D voxels. The 2D image features correspond to the visible region within the voxel, and these features are relatively reliable. However, the number of voxels in the invisible region is significantly greater than those in the visible region. Our horizontal window-to-global attention (W2G) module first diffuses the features from the visible region horizontally and then globally, thereby enhancing the retention of effective features.

Another constraint of prior approaches lies in their disregard for the variation in the importance of voxels at different positions. From a human perceptual standpoint, the features in transition regions between different objects are more pronounced. To address this objective, we introduce the notion of Internal-External indicators (IoE) to quantitatively ascertain the significance of a given position along the horizontal and vertical dimensions. Divergent from the Local Geometric Anisotropy (LGA) (Li et al. 2019b), our Internal-External indicators (IoE) prioritize the transitional changes in position rather than the position itself, and they are comparatively more straightforward to implement. And we further introduce the Internal-External Position Awareness Loss (IoE-PALoss).

Our contributions can be summarized as follows:

- We propose a novel H2GFormer to leverage the RGB data for semantic scene completion. H2GFormer effectively utilizes 2D features through a progressive feature reconstruction process across various directions.
- We introduce horizontal window-to-global attention (W2G) module to enhance the model's focus on variations of voxels along the horizontal direction.
- We introduce a novel Internal-External Position Awareness Loss (IoE-PALoss), which highlights voxels in transition between different directions of objects and alleviates the excessive focus on redundant internal object information.
- H2GFormer achieves a performance improvement of 19.7% over the state-of-the-art VoxFormer-S (Li et al. 2023) and 8.9% over VoxFormer-T on the SemanticKITTI (Behley et al. 2019) dataset.

Related Work

3D Semantic Scene Completion

SSCNet (Song et al. 2017) was the first to introduce the task of 3D semantic scene completion, considering both scene completion and semantic labeling jointly. The initial works (Li et al. 2019a; Chen, Garbade, and Gall 2019; Wang et al. 2019; Chen, Xing, and Zeng 2020; Li et al. 2020a; Chen et al. 2020; Cai et al. 2021; Dourado et al. 2021) mostly relied on 3D CNN volume networks, used for processing relatively smaller-scale indoor datasets like NYUv2 (Silberman et al. 2012) and NYUCAD (Firman et al. 2016). Recently, with the release of large-scale datasets such as SemanticKITTI (Behley et al. 2019), there has been a growing interest in outdoor semantic scene completion tasks (SSC). Some research studies employ LiDAR data as input for scene completion (Roldao, de Charette, and Verroust-Blondet 2020; Xia et al. 2023; Yang et al. 2021). However, LiDAR sensors can be costly. Hence, other research is continuously exploring the possibility of utilizing only 2D images as input for scene completion.

MonoScene (Cao and de Charette 2022) utilizes projection to model 2D images into 3D. TPVFormer (Huang et al. 2023), OccFormer (Zhang, Zhu, and Du 2023) and SurroundOcc (Wei et al. 2023) have explored the use of multi-view semantic information such as tri-perspective view (TPV) and bird's eye view (BEV) to enhance SSC tasks. OccDepth (Miao et al. 2023) introduces a stereo allocation module to enhance the fusion of depth perception. VoxFormer (Li et al. 2023) designs transformer-based networks to more effectively elevate 2D features to 3D.

Compared to the previous works, our camera-based perception network focuses on the positional importance of voxels within the scene, enabling a more effective feature extraction and diffusion.

Loss Function for 3D Dense Prediction

In outdoor scenes, 3D SSC differs significantly from 2D segmentation. The importance of voxels at different directions and positions varies significantly for 3D SSC. And voxel counts in transition regions between objects are smaller than within object interiors or exteriors. Thus, selecting an appropriate loss function is crucial for effective and accurate network training in SSC. Currently, various classical loss functions are available for this purpose.

1) Weighted Cross-Entropy Loss: The weighted cross-entropy loss introduces class-specific weighting factors $\omega_c \in [0, 1]$ upon the foundation of cross-entropy loss. This approach emphasizes the significance of less-sampled categories, thereby addressing class imbalance to some extent. And the weighting parameters can be manually configured.

2) Scene-Class Affinity Loss: (Cao and de Charette 2022) introduces the Scene-Class Affinity Loss, which is aimed at the simultaneous optimization of class-wise derivable precision, recall, and specificity metrics.

3) Position Aware loss: The Position Aware Loss introduced by (Li et al. 2019b) employs the Local Geometric Anisotropy factor to amplify the response toward voxels containing intricate details.

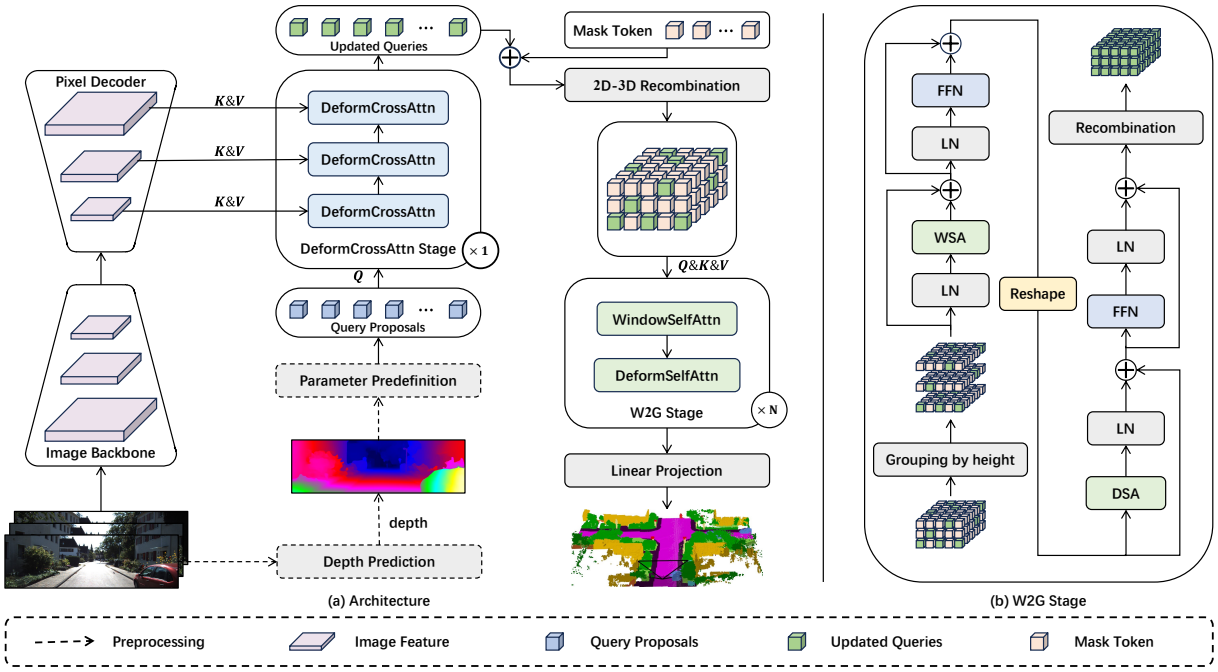

Figure 2: (a) The overall architecture of H2GFormer. The image backbone and pixel decoder extract features from single or multiple images, which are then fed into the DeformCrossAttn Stage using a hierarchical multi-scale feature allocation strategy. Subsequently, the predefined features of the visible region are combined with a mask token and propagated through the W2G stage to diffuse the features. Finally, upsampling and projection are performed to obtain the semantic scene completion results. (b) The structure of the W2G stage, which involves windowed self-attention (WSA) and deformable self-attention (DSA).

Methodology

In this section, we present the architecture of H2GFormer. H2GFormer predicts the occupancy and semantic category of each 3D voxel from 2D images for SSC based on the Transformer. Each 3D voxel will be mapped to one of (N+1) semantic labels $C = \{c_0, c_1, ..., c_N\}$, where c_0 represents an empty voxel, and N is the number of semantic categories. Specifically, as illustrated in Figure 2, we input the 2D images into ResNet50 to extract features. Then, the RGB features are fused through the pixel-decoder module and hierarchically fed into the Deformable Cross-Attention module (Zhu et al. 2020). By utilizing the Deformable Cross-Attention module, we are enable to transform 2D image features into 3D visible region features. Subsequently, the horizontal window-to-global attention (W2G) module diffuses the voxel with Mask Token from a horizontal to a global sequence, covering the entire voxel representation. Finally, the features are processed through linear projection to obtain the dense semantic prediction map $Y_t \in \mathbb{R}^{H \times W \times Z \times (N+1)}$ for voxel.

Below, we introduce the details of H2GFormer from the following aspects: 1) Predefined Parameters, 2) Pixel Decoder, 3) Deformable Cross-Attention module, 4) Horizontal Window-to-Global Attention module, 5) Loss Function.

Predefined Parameters

Query Proposal. The voxels in the three-dimensional space corresponding to the two-dimensional image features are sparse. As a result, the feature extraction for visible-occupied voxels is more reliable. Following the same approach as the Class-Agnostic Query Proposal Stage in (Li et al. 2023), we define a total of N_p voxel queries $Q_p \in \mathbb{R}^{N_p \times d}$, where d denotes the feature dimension.

Mask Token. Opting to extract features solely for visible-occupied voxels from the image features implies that the features for the remaining voxels need to be initialized using another set of learnable parameters. Similar to (Li et al. 2023; He et al. 2022), we introduce mask token $m \in \mathbb{R}^d$, to indicate the learnable vector representing the missing voxel features that need to be predicted based on the features of visible voxels.

Pixel Decoder

High-resolution features can enhance the predictive performance of the model (Ronneberger, Fischer, and Brox 2015; Wang et al. 2020; Zou et al. 2021), particularly concerning small-sized objects. Experiments conducted in VoxFormer demonstrate that directly inputting high-resolution feature maps into the Cross-Attention module yields superior performance compared to utilizing multi-scale feature maps. However, we consider that due to the sparse mapping of information from 2D images to the 3D real-world space, a step-wise extraction of 3D features from 2D image features is necessary. This design entails initially extracting the overall information from the image before capturing localized details. Inspired by Mask2Former (Cheng, Schwing, and

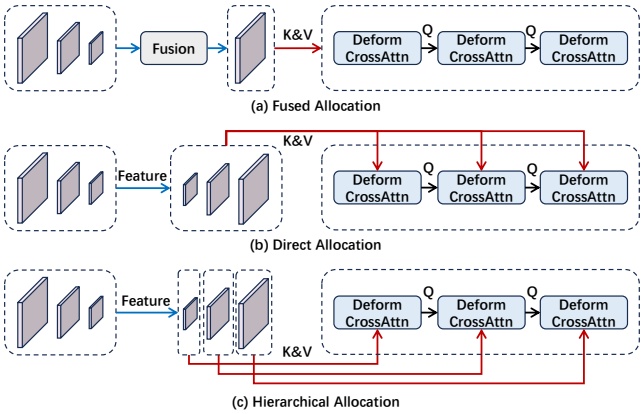

Figure 3: (a) The fused allocation; (b) The direct allocation; (c) Our hierarchical multi-scale feature allocation strategy.

Kirillov 2021; Cheng et al. 2022), we have devised a pixel decoder that receives image features from the backbone network. It incrementally upsamples low-resolution features to acquire high-resolution pixel features. Subsequently, these features of varying resolution scales are sequentially fed into the Cross-Attention module, following an ascending order from low to high resolution. Specifically, the structural distinction between this design and the direct utilization of high-resolution features is depicted in Figure 3. Figure 3(a) illustrates early fusion followed by feature extraction using a single-resolution feature map. Figure 3(b) illustrates the repeated utilization of multi-level feature maps. Figure 3(c) illustrates our designed structure capable of hierarchical feature extraction, thus capturing more intricate details.

Deformable Cross-Attention Module

Upon obtaining feature maps from the pixel decoder, we utilize query proposals Q_p to learn the 3D features from image features. Here, we employ deformable attention (Zhu et al. 2020), which selectively samples N_s points around the reference point for attention computation, thereby enhancing efficiency. Mathematically, deformable attention(DA) update the query q using the following equation:

$$DeformAttn(q, p, F) = \sum_{s=1}^{N_s} A_s W_s F(p + \delta p_s) \quad (1)$$

where p represents the reference point, F denotes the input feature, N_s indicates the number of sampled points. W_s and A_s are learnable weights, and $F(p+\delta p_s)$ represents the feature of sampled point locations extracted through bilinear interpolation.

Specifically, within the deformable cross-attention module, drawing inspiration from (Li et al. 2023), we employ the same settings to select reference points p for each query proposal q_p. In the case of utilizing multiple frames of view, we perform a weighted summation of the features sampled on views where selected reference points exist, resulting in the output of the deformable cross-attention.

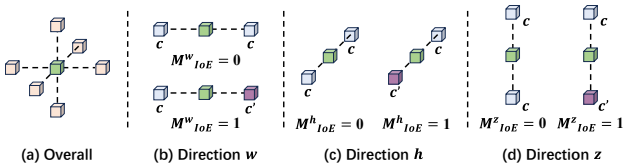

Figure 4: Internal-External indicator: indicate the positional significance of object transition edge voxels and object internal voxels. (a) Current voxel and surrounding voxels; (b)-(d) Values computed for IoE along the w, h, and z directions.

Horizontal Window-to-Global Attention Module

After several layers of deformable cross-attention modules, the query proposals extract 3D features of the visible regions from the 2D features and are updated as $\hat{Q}_p$. Subsequently, we integrate the predefined mask tokens of the invisible regions with the updated query proposals to obtain the complete voxel features, which serve as the initial input $F^{3D} \in \mathbb{R}^{h\times w\times z\times d}$ for the Horizontal Window-to-Global Attention module. This module comprises two steps. First, for the input 3D features, semantic information is aggregated along the horizontal direction. Subsequently, the semantic information is propagated to global voxels. In specific terms, we merge the height dimension into the batch dimension, resulting in the feature $F^{3D-H} \in \mathbb{R}^{z\times h\times w\times d}$ and apply windowed self-attention (Liu et al. 2021) to each height-level feature map. This step can be formalized as follows:

$$F^{3D-H} = WMSA(F^{3D-H}) \quad (2)$$

where WMSA is multi-head windowed self attention modules.

Subsequently, we restore the reshaped 3D features back to their original shape $F^{3D} \in \mathbb{R}^{h\times w\times z\times d}$ and further refine them using deformable self-attention, resulting in the feature tensor $\hat{F}^{3D} \in \mathbb{R}^{h\times w\times z\times d}$:

$$DSA(F^{3D}, F^{3D}) = DA(f, p, F^{3D}) \quad (3)$$

where f represents the updated query proposal located at the reference point p.

Loss Function

Given the sparse arrangement of objects between outdoor scenes, compared to indoor scenarios, voxels at object edges and transitions hold higher positional significance. To address this, drawing inspiration from PALoss (Li et al. 2019b), we introduce a variant that measures whether a voxel is located at an object edge. To quantify the Internal-External indicators (IoE) for specific voxels, we focus on the semantic categories of voxels located around the current voxel along the three coordinate axes. Specifically, for each direction, if the semantic categories of the voxels before and after the current voxel are different, we consider it as an edge voxel, as illustrated in Figure 4. Given a voxel p, its IoE is calculated based on six neighboring voxels along the three directions and can be expressed as the following formula:

$$M_{IoE}(p) = \sum_{i=1}^{3} (c^i_{p-1} \oplus c^i_{p+1}) \quad (4)$$

Methods	H2GFormer-T (Ours)			H2GFormer-S (Ours)			VoxFormer-T			VoxFormer-S			MonoScene		
Range	**12.8m**	**25.6m**	**51.2m**	**12.8m**	**25.6m**	**51.2m**	**12.8m**	**25.6m**	**51.2m**	**12.8m**	**25.6m**	**51.2m**	**12.8m**	**25.6m**	**51.2m**
IoU (%)	**67.49**	58.51	**44.69**	66.42	**58.61**	44.57	65.38	57.69	44.15	65.35	57.54	44.02	38.42	38.55	36.80
Precision (%)	**82.00**	73.98	62.26	79.02	**74.24**	62.17	76.54	69.95	62.06	77.65	70.85	**62.32**	51.22	51.96	52.19
Recall (%)	79.23	73.66	**61.29**	80.64	73.58	61.16	**81.77**	**76.70**	60.47	80.49	75.39	59.99	60.60	59.91	55.50
mIoU	**23.43**	**20.37**	**14.29**	20.49	18.39	13.73	21.55	18.42	13.35	17.66	16.48	12.35	12.25	12.22	11.30
car	**48.81**	**40.15**	**28.21**	45.07	39.30	27.60	44.90	37.46	26.54	39.78	35.24	25.79	24.34	24.64	23.29
bicycle	1.76	1.73	0.95	0.59	1.86	0.50	**5.22**	**2.87**	**1.28**	3.04	1.48	0.59	0.07	0.23	0.28
motorcycle	**9.75**	2.27	**0.91**	3.61	**2.40**	0.47	2.98	1.24	0.56	2.84	1.10	0.51	0.05	0.20	0.59
truck	10.29	**15.06**	6.80	15.24	9.34	**10.00**	9.80	10.38	7.26	7.50	7.47	5.63	**15.44**	13.84	9.29
other-veh.	16.54	**10.71**	9.32	16.22	7.71	7.39	**17.21**	10.61	7.81	8.71	4.98	3.77	1.18	2.13	2.63
person	2.00	2.75	1.15	2.00	3.06	1.54	**4.44**	**3.50**	1.93	4.10	3.31	1.78	0.90	1.37	**2.00**
bicyclist	0.52	0.86	0.10	2.78	3.89	2.88	2.65	3.92	1.97	**6.82**	**7.14**	3.32	0.54	1.00	1.07
motorcyclist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
road	**80.17**	**71.72**	**57.00**	76.83	70.79	56.08	75.45	66.15	53.57	72.40	65.74	54.76	57.37	57.11	55.89
parking	**27.90**	**28.42**	**21.74**	18.96	19.03	17.83	21.01	23.96	19.69	10.79	18.49	15.50	20.04	18.60	14.75
sidewalk	**55.61**	**41.18**	**29.37**	49.72	39.68	29.12	45.39	34.53	26.52	39.35	33.20	26.35	27.81	27.58	26.50
other-grnd	0.00	0.50	0.34	0.01	0.72	0.45	0.00	0.76	0.42	0.00	1.54	0.70	**1.73**	**2.00**	**1.63**
building	**26.66**	**33.12**	**20.51**	21.83	28.80	19.74	25.13	29.45	19.54	17.91	24.09	17.65	16.67	15.97	13.55
fence	**17.96**	**12.27**	**7.98**	14.88	11.30	7.24	16.17	11.15	7.31	12.98	10.63	7.64	7.57	7.37	6.60
vegetation	**46.53**	**40.34**	**27.44**	43.00	37.51	26.25	43.55	38.07	26.10	40.50	34.68	24.39	19.52	19.68	17.98
trunk	21.37	**15.18**	**7.80**	13.38	12.16	6.80	**21.39**	12.75	6.10	15.81	10.64	5.08	2.02	2.57	2.44
terrain	**48.46**	**44.66**	**36.26**	43.45	41.75	34.42	42.82	39.61	33.06	32.25	35.08	29.96	31.72	31.59	29.84
pole	**21.46**	**16.91**	**9.88**	15.10	13.12	7.88	20.66	15.56	9.15	14.47	11.95	7.11	3.10	3.79	3.91
traf.-sign	9.33	**9.16**	**5.81**	6.45	7.06	4.68	**10.63**	8.09	4.94	6.19	6.29	4.18	3.69	2.54	2.43

Table 1: Quantitative comparisons against the camera-based baseline methods on SemanticKITTI validation set. We compare performance across three volume ranges, from short-range to long-range regions. The performance of the top two models is indicated using bold and underlined formatting.

where c^i_{p-1} represents the semantic label of the voxel before the current voxel along direction i, c^i_{p+1} represents the semantic label of the voxel after the current voxel along direction i. If voxel c^i_{p-1} and c^i_{p+1} have the same semantic label, then $c^i_{p-1} \oplus c^i_{p+1} = 0$, otherwise $c^i_{p-1} \oplus c^i_{p+1} = 1$.

Based on IoE values, IoE PA-Loss is defined as follows:

$$L_{IoE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=c_0}^{c_M} \alpha M^n_{IoE}\hat{y}_{nc} log(\frac{e^{y_{nc}}}{\sum_c e^{y_{nc}}}) \quad (5)$$

where N is the total number of voxels used for computing this loss, c indexes class, α represents the scaling coefficient and M^n_{IoE} is the IoE of voxels n. $\hat{y}_{nc}$, y_{nc} are the one-hot vector of the ground truth labels and the corresponding predictions in class c.

We also employed the scene-class affinity loss (Cao and de Charette 2022) and the universal weighted cross-entropy loss (Roldao, de Charette, and Verroust-Blondet 2020). And the overall loss is the sum of these terms.

Experiments

Experimental Setup

Dataset. We evaluate the performance of H2GFormer on the SemanticKITTI (Behley et al. 2019) dataset, which comprises densely annotated semantic labels for each LiDAR scan from 22 outdoor driving scenes of the KITTI Odometry Benchmark (Geiger, Lenz, and Urtasun 2012). This dataset voxelizes point clouds and labels the entire scene within a size of 51.2m×51.2m×6.4m, resulting in a voxel grid of dimensions 256×256×32. Furthermore, the voxel grid annotations encompass a total of 20 categories, which include 19 semantic classes and 1 empty class. Regarding the sparse input to an SSC model, SemanticKITTI provides RGB images and LiDAR scans. In this study, we investigated image-based SSC similar to (Li et al. 2023).

Implementation Details. In the predefined portion of the Query proposal for deformable cross-attention, we directly utilize the same approach as in (Li et al. 2023) to obtain the initial query proposal for the visible region. In the image backbone section, we employ ResNet50 (He et al. 2016) to process the input RGB images of size 1220×370 and extract image features. In the pixel decoder module, we directly utilize FPN (Lin et al. 2017) to process image features and obtain three feature maps of sizes 1/4, 1/8, and 1/16 of the input image. Subsequently, these feature maps are sequentially inputted into a concatenated sequence of three deformable cross-attention modules. The feature dimension is set as $d = 128$. The number of cross-attention modules is 3, and the number of W2A modules is 2. The W2A module alternates between using windowed self-attention with regular window and shifted window (Liu et al. 2021). We trained the model for 20 epochs with a learning rate of 2×10^{-4}. Similar to (Li et al. 2023), we also provided two versions of H2GFormer, one that takes only the current image as input (H2GFormer-S), and another that takes the current image along with the previous 4 images as input (H2GFormer-T).

Methods	IoU	mIoU	car	bicycle	motorcycle	truck	other-veh.	person	bicyclist	motorcyclist	road	parking	sidewalk	other-grnd	building	fence	vegetation	trunk	terrain	pole	traf.-sign
LMSCNet*	31.38	7.07	14.30	0.00	0.00	0.30	0.00	0.00	0.00	0.00	46.70	13.50	19.50	3.10	10.30	5.40	10.80	0.00	10.40	0.00	0.00
3DSketch*	26.85	6.23	17.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.70	0.00	19.80	0.00	12.10	3.40	12.10	0.00	16.10	0.00	0.00
AICNet*	23.93	7.09	15.30	0.00	0.00	0.70	0.00	0.00	0.00	0.00	39.30	19.80	18.30	1.60	9.60	5.00	9.60	1.90	13.50	0.10	0.00
JS3C-Net*	34.00	8.97	20.10	0.00	0.00	0.80	4.10	0.00	0.20	0.20	47.30	19.90	21.70	2.80	12.70	8.70	14.20	3.10	12.40	1.90	0.30
MonoScene	34.16	11.08	18.80	0.50	0.70	3.30	4.40	1.00	1.40	0.40	54.70	24.80	27.10	5.70	14.40	11.10	14.90	2.40	19.50	3.30	2.10
SurroundOcc	34.72	11.86	20.60	1.60	1.20	1.40	4.40	1.40	2.00	0.10	56.90	30.20	28.30	6.80	15.20	11.30	14.90	3.40	19.30	3.90	2.40
Symphonies	41.07	13.02	22.10	1.70	1.30	1.90	**5.80**	**2.20**	1.30	**0.50**	55.70	25.30	26.80	4.90	21.30	13.10	22.90	8.20	19.50	6.80	5.80
TPVFormer	34.25	11.26	19.20	1.00	0.50	3.70	2.30	1.10	2.40	0.30	55.10	27.40	27.20	6.50	14.80	11.00	13.90	2.60	20.40	2.90	1.50
OccFormer	34.53	12.32	21.60	1.50	**1.70**	1.20	3.20	**2.20**	1.10	0.20	55.90	**31.50**	30.30	6.50	15.70	11.90	16.80	3.90	21.30	3.80	3.70
VoxFormer-S	42.95	12.20	20.80	1.00	0.70	3.50	3.70	1.40	**2.60**	0.20	53.90	21.10	25.30	5.60	19.80	11.10	22.40	7.50	21.30	5.10	4.90
VoxFormer-T	43.21	13.41	21.70	**1.90**	1.60	3.60	4.10	1.60	1.10	0.00	54.10	25.10	26.90	**7.30**	23.50	13.10	24.40	8.10	24.20	6.60	5.70
H2GFormer-S (Ours)	**44.20**	13.72	23.40	0.80	0.90	4.80	4.10	1.20	2.50	0.10	56.40	26.50	28.60	4.90	22.80	13.30	24.60	9.10	23.80	6.40	6.30
H2GFormer-T (Ours)	43.52	**14.60**	**23.70**	0.60	1.20	**5.20**	5.00	1.10	0.10	0.00	**57.90**	30.00	**30.40**	6.90	**24.00**	**14.60**	**25.20**	**10.70**	**25.80**	**7.50**	**7.10**

Table 2: Quantitative comparisons with the state-of-the-art methods on SemanticKITTI hidden test set. * indicates the results obtained from the official code of these methods when using RGB inputs, reported in MonoScene.

Evaluation Metrics. We employed Intersection over Union (IoU) and mean IoU (mIoU) metrics to evaluate the SSC task. IoU metric evaluates scene completion quality by examining occupancy prediction maps. In SSC, mIoU more effectively gauges the model's scene understanding ability, enhancing its applicability to advanced tasks. Additionally, we evaluated the IoU and mIoU within volumes of sizes 12.8m×12.8m×6.4m, 25.6m×25.6m×6.4m, and 51.2m×51.2m×6.4m.

Comparison against Baseline Methods

Our H2GFormer maintains consistency with VoxFormer (Li et al. 2023) in terms of depth estimation (Shamsafar et al. 2022), thereby validating the superior performance of our approach in aggregating 3D semantic features. Therefore, in Table 1, we compare H2GFormer with recent camera-based VoxFormer (Li et al. 2023) and MonoScene (Cao and de Charette 2022) on the SemanticKITTI validation dataset. We evaluate and compare the performance of these various baseline methods in both short-range and long-range regions. Our H2GFormer consistently outperforms VoxFormer inside three volumes (51.2m, 25.6m, and 12.8m). Specifically, H2GFormer-S achieves improvements of mIoU of 11.17%, 11.59%, and 16.02% over VoxFormer-S in the three volumes, respectively. Furthermore, our approach demonstrates excellent performance in classes that exhibit horizontal scene layouts, such as road, terrain, sidewalk, parking, and building.

Comparisons with the State-of-the-Art Methods

In Table 2, we compare our proposed method with state-of-the-art camera-based approaches on SemanticKITTI hidden test set, including MonoScene (Cao and de Charette 2022), SurroundOcc (Wei et al. 2023), Symphonies (Jiang et al. 2023), TPVFomer (Huang et al. 2023), OccFormer (Zhang, Zhu, and Du 2023), and VoxFormer (Li et al. 2023). Our H2GFormer achieves the highest mIoU (14.60), surpassing the previous highest of 13.41 (Li et al. 2023) by an 8.87% improvement.

Methods	Spatial resolution $\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	IoU	mIoU
Original FPN	✓				44.29	10.67
Original FPN		✓			44.19	11.85
Original FPN			✓		44.47	12.06
Original FPN				✓	44.30	11.85
Original FPN		✓	✓	✓	44.34	11.98
Original FPN	✓	✓	✓		44.23	12.13
Original FPN	✓	✓	✓	✓	44.01	12.23
Pixel Decoder		✓	✓	✓	44.19	12.02
Pixel Decoder	✓	✓	✓		44.20	**12.36**

Table 3: Ablation study for pixel decoder. The pixel decoder utilizes our hierarchical multi-scale feature allocation strategy. Spatial resolution is relative to the input image size.

Ablation Study

In this section, we conduct ablation studies on the SemanticKITTI validation set to validate the effectiveness of the aforementioned components.

Ablation on the Pixel Decoder. The ablation study for the pixel decoder component is presented in Table 3. The three sections in Table 3 correspond to the three structures in Figure 3 from top to bottom. We primarily compared the performance difference between the proposed hierarchical output multi-scale feature map Pixel decoder and the original FPN. We observe that the use of the original FPN does not effectively leverage the multi-scale resolution features, whereas our proposed hierarchical multi-scale strategy ef-

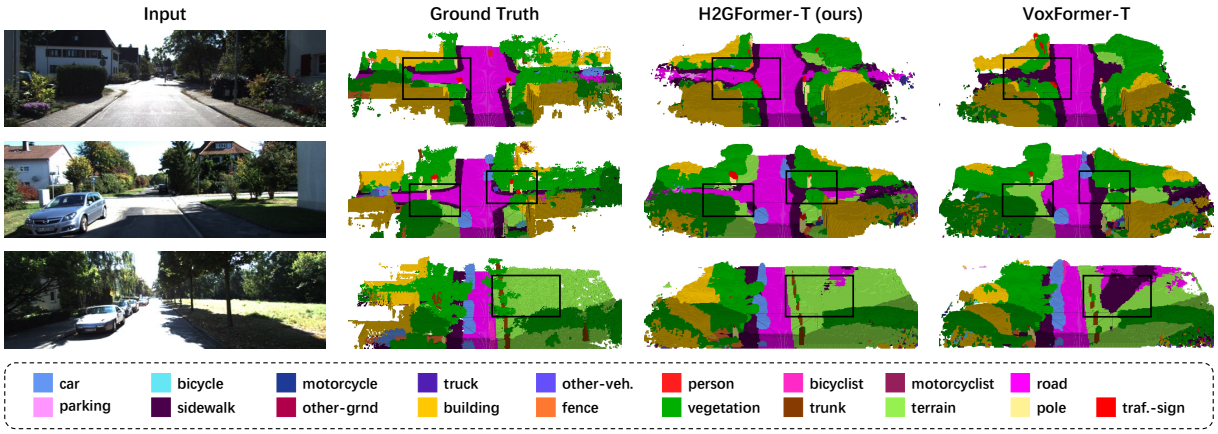

Figure 5: Qualitative results of our method and VoxFormer (Li et al. 2023) on SemanticKITTI validation set. H2GFormer demonstrates enhanced capability in distinguishing horizontal variations, such as road, terrain, and sidewalk. Simultaneously, H2GFormer also excels in completing small objects like traffic signs.

W2G module 1		W2G module 2		Layer	IoU	mIoU
-	DSA	-	DSA	2	44.34	12.61
WA	DSA	-	DSA	2	44.68	12.75
-	DSA	SWA	DSA	2	44.72	12.95
DSA	DSA	DSA	DSA	2	43.89	12.71
WA	DSA	WA	DSA	2	44.92	13.38
SWA	DSA	SWA	DSA	2	45.14	13.32
WA	DSA	SWA	DSA	2	44.57	**13.73**
WA	DSA	SWA	DSA	4	44.86	13.58

Table 4: Ablation study for W2G module. DSA, WA, and SWA respectively indicate deformable self-attention, window attention, and shifted window attention. Layer indicates the number of W2G modules.

Methods	IoU (%)	mIoU (%)
Ours-S	44.57	**13.73**
Ours-S w/o pixel decoder	44.63	13.42
Ours-S w/o W2G module	44.34	12.61
Ours-S w/o IoEPALoss	44.74	13.38
VoxFormer-S (Li et al. 2023)	44.02	12.35
Ours-T	44.69	**14.29**
Ours-T w/o pixel decoder	44.90	14.19
Ours-T w/o W2G module	44.43	13.69
Ours-T w/o IoEPALoss	44.76	13.95
VoxFormer-T (Li et al. 2023)	44.15	13.35

Table 5: Ablation study for architecture.

fectively utilizes features of different resolutions. Furthermore, employing feature maps at 1/4, 1/8, and 1/16 resolutions achieves higher mIoU values.

Ablation on the Horizontal Window-to-Global Attention Module. The ablation study of the W2G module is presented in Table 4. We conducted a comparative analysis of the performance differences between directly using deformable self-attention, using window attention, and using shifted window attention during the process of feature propagation in the horizontal direction. We observe that alternating between the two types of window attention leads to better performance.

Ablation on the Architecture. The ablation study of the architecture is presented in Table 5. We observe that both in H2GFormer-S and H2GFormer-T, the pixel decoder, the Horizontal Window-to-Global Attention module, and the Internal-External Position Awareness Loss contribute to the best results. Furthermore, H2GFormer-T achieved a performance improvement of 15.7% (+0.67 IoU, +1.94 mIoU) over the baseline VoxFormer-S (Li et al. 2023) and a 7% (+0.54 IoU, +0.94 mIoU) improvement over VoxFormer-T.

Qualitative Visualizations

The qualitative comparison of our H2GFormer-T and VoxFormer-T on the SemanticKITTI validation set is visually illustrated in Figure 5. As evident from the black rectangular boxes, H2GFormer demonstrates a stronger capability in distinguishing horizontal variations such as roads, terrain, and sidewalks. Additionally, H2GFormer excels in completing small objects like traffic signs.

Conclusion

In this paper, we present H2GFormer, a powerful and efficient camera-based 3D semantic scene completion framework. To better utilize image features of different resolutions, we proposed a hierarchical multi-scale feature allocation strategy. Additionally, we focused on the variations of the importance of voxels at different spatial positions in 3D space through Horizontal Window-to-Global Attention and Internal-External Position Awareness Loss, achieving improved geometric and semantic performance. Experimental results show that our method contributes to the understanding of semantic information on the horizontal plane and achieves excellent performance on the SemanticKITTI dataset.

Acknowledgments

This study is partially supported by National Natural Science Foundation of China (62176016, 72274127), National Key R&D Program of China (No. 2021YFB2104800), Guizhou Province Science and Technology Project: Research and Demonstration of Sci. & Tech Big Data Mining Technology Based on Knowledge Graph (supported by Qiankehe[2021] General 382), and Capital Health Development Research Project(2022-2-2013).

References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.

Cai, Y.; Chen, X.; Zhang, C.; Lin, K.-Y.; Wang, X.; and Li, H. 2021. Semantic scene completion via integrating instances and scene in-the-loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 324–333.

Cao, A.-Q.; and de Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.

Chen, X.; Lin, K.-Y.; Qian, C.; Zeng, G.; and Li, H. 2020. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4193–4202.

Chen, X.; Xing, Y.; and Zeng, G. 2020. Real-time semantic scene completion via feature aggregation and conditioned prediction. In *2020 IEEE International Conference on Image Processing (ICIP)*, 2830–2834. IEEE.

Chen, Y.-T.; Garbade, M.; and Gall, J. 2019. 3d semantic scene completion from a single depth image using adversarial training. In *2019 IEEE International Conference on Image Processing (ICIP)*, 1835–1839. IEEE.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Cheng, R.; Agia, C.; Ren, Y.; Li, X.; and Bingbing, L. 2021. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, 2148–2161. PMLR.

Dourado, A.; De Campos, T. E.; Kim, H.; and Hilton, A. 2021. EdgeNet: Semantic scene completion from a single RGB-D image. In *2020 25th international conference on pattern recognition (ICPR)*, 503–510. IEEE.

Firman, M.; Mac Aodha, O.; Julier, S.; and Brostow, G. J. 2016. Structured prediction of unobserved voxels from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5431–5440.

Garbade, M.; Chen, Y.-T.; Sawatzky, J.; and Gall, J. 2019. Two stream 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9223–9232.

Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Liu, W.; and Wang, X. 2023. Symphonize 3D Semantic Scene Completion with Contextual Instance Queries. *arXiv preprint arXiv:2306.15670*.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020a. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3351–3359.

Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; and Reid, I. 2019a. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7693–7702.

Li, J.; Liu, Y.; Yuan, X.; Zhao, C.; Siegwart, R.; Reid, I.; and Cadena, C. 2019b. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters*, 5(1): 219–226.

Li, S.; Zou, C.; Li, Y.; Zhao, X.; and Gao, Y. 2020b. Attention-based multi-modal fusion network for semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11402–11409.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9087–9098.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; and Zhou, S. 2023. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*.

Roldao, L.; de Charette, R.; and Verroust-Blondet, A. 2020. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, 111–119. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shamsafar, F.; Woerz, S.; Rahim, R.; and Zell, A. 2022. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2417–2426.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. 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*, 746–760. Springer.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1746–1754.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, Y.; Tan, D. J.; Navab, N.; and Tombari, F. 2019. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8608–8617.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*.

Xia, Z.; Liu, Y.; Li, X.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; and Qiao, Y. 2023. SCPNet: Semantic Scene Completion on Point Cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17642–17651.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3101–3109.

Yang, X.; Zou, H.; Kong, X.; Huang, T.; Liu, Y.; Li, W.; Wen, F.; and Zhang, H. 2021. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3555–3562. IEEE.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. *arXiv preprint arXiv:2304.05316*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zou, H.; Yang, X.; Huang, T.; Zhang, C.; Liu, Y.; Li, W.; Wen, F.; and Zhang, H. 2021. Up-to-Down Network: Fusing Multi-Scale Context for 3D Semantic Scene Completion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 16–23. IEEE.