Symphonize 3D Semantic Scene Completion with Contextual Instance Queries

Haoyi Jiang[1*] Tianheng Cheng[1*] Naiyu Gao[2] Haoyang Zhang[2] Tianwei Lin[2]
Wenyu Liu[1] Xinggang Wang [1†]
[1] School of EIC, Huazhong University of Science & Technology
[2] Horizon Robotics
{haoyi_jiang,thch,liuwy,xgwang}@hust.edu.cn
{naiyu.gao,haoyang.zhang,tianwei.lin}@horizon.cc

Abstract

3D Semantic Scene Completion (SSC) has emerged as a nascent and pivotal undertaking in autonomous driving, aiming to predict the voxel occupancy within volumetric scenes. However, prevailing methodologies primarily focus on voxel-wise feature aggregation, while neglecting instance semantics and scene context. In this paper, we present a novel paradigm termed ***Symphonies (Scene-from-Insts)****, that delves into the integration of instance queries to orchestrate 2D-to-3D reconstruction and 3D scene modeling. Leveraging our proposed Serial Instance-Propagated Attentions, Symphonies dynamically encodes instance-centric semantics, facilitating intricate interactions between the image and volumetric domains. Simultaneously, Symphonies fosters holistic scene comprehension by capturing context through the efficient fusion of instance queries, alleviating geometric ambiguities such as occlusion and perspective errors through contextual scene reasoning. Experimental results demonstrate that Symphonies achieves state-of-the-art performance on the challenging SemanticKITTI and SSCBench-KITTI-360 benchmarks, yielding remarkable mIoU scores of 15.04 and 18.58, respectively. These results showcase the promising advancements of our paradigm. The code for our method is available at* `https://github.com/hustvl/Symphonies`.

1. Introduction

The advent of autonomous driving has brought forth novel challenges in the realm of 3D perception. In the pursuit of safe navigation and obstacle avoidance, autonomous vehicles must be equipped with the ability to accurately predict the occupancy of their immediate surroundings. This task, however, is not a facile endeavor, given the inherent complexities of the real world, characterized by clutter, ambiguity, and rapid evolution.

3D Semantic Scene Completion (SSC) formulates this challenge as the reconstruction of occupancy and semantics for every voxel grid within a 3D scene. Recent advancements in vision-based solutions, such as MonoScene [3] and OccDepth [30], adopt 3D convolutional networks to elevate 2D image features to 3D volumes. TPVFormer [16], OccFormer [46], and CTF-Occ [38] explore decomposing 3D volumes into multiple coarse view representations and enhancing voxel interactions using Transformer [23, 39, 48] architectures.

Despite these advancements, contemporary approaches tend to prioritize voxel-wise modeling for 3D scenes resorting to pixel-voxel projection [6, 23, 31, 32] for the feature promotion from 2D to 3D. While focusing on these localized representations, they inadvertently neglect higher-level instance semantics, leading to vulnerability to geometric ambiguities arising from occlusion and perspective errors. Humans, in contrast, naturally perceive and comprehend through the concept of "instance", rather than isolated pixels or voxels, each imbued with semantic significance and cohesively contributing to the contextual whole of a scene. In light of these limitations, a fundamental question arises: ***How can we leverage the notion of instances to steer 3D scene modeling and 2D-to-3D reconstruction?***

Drawing inspiration from this notion, we propose Symphonies (Scene-from-Insts), a novel method that leverages contextual instance queries derived from image inputs to enhance scene modeling, exploiting inherent instance semantics and scene context. Stemming from this basis, we propose Serial Instance-Propagated Attentions to intricately interact with image and voxel features, deformably aggregating instance-centric semantics. This seamless interaction bridges the gap between low-level pixel or voxel representations and high-level semantics, facilitating feature promo-

*Equal contribution. Work done during Haoyi Jiang's internship at Horizon Robotics.
†Corresponding author.

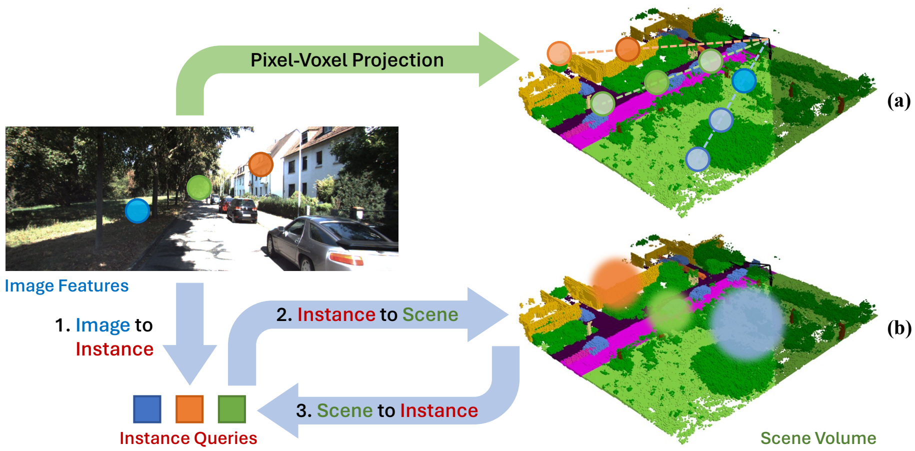

Figure 1. **Comparison between voxel-wise modeling (a) and Symphonies (b).** Conventional methods primarily depend on Inverse Perspective Mapping (IPM)-based voxel-pixel projection and voxel-wise feature aggregation, resulting in geometric ambiguities and computational redundancy. In contrast, Symphonies leverages instance queries as intermediaries to engage with image and voxel features, thus exploiting instance semantics and enhancing the contextual comprehension of the scene.

tion and scene modeling, as illustrated in Fig. 1. Furthermore, the synergistic fusion of multiple instance queries collectively enriches broader contextual information necessary for robust scene reasoning, contributing to the alleviation of geometric ambiguities. In tandem, we introduce the Depth-Rectified Voxel Proposal Layer to refine the initial geometry, elevating 2D image features to the implicit surface of the scene.

To evaluate the effectiveness of our method, extensive experiments are conducted on SemanticKITTI [1] and SSCBench-KITTI-360 [21, 24] datasets. Symphonies achieves a remarkable state-of-the-art performance of 15.04 and 18.58 mIoU, respectively, significantly outperforming previous vision-based methods by a substantial margin. Ablation experiments further underscore the promising advancements of our approach in the field of SSC. In summary, our contributions involve:

- We introduce Symphonies, a pioneering paradigm for 3D Semantic Scene Completion (SSC), which delves into modeling instance-centric semantics with our proposed Serial Instance-Propagated Attentions, facilitating interactions between image features and scene modeling.
- Symphonies captures global context of scenes through the fusion of instance queries, thereby alleviating geometric ambiguities. To further enhance robustness, we introduce the Depth-Rectified Voxel Proposal Layer which explicitly refines the initial geometry with estimated depth information.
- Our proposed method achieves state-of-the-art performance on the SemanticKITTI and SSCBench-KITTI-360 benchmarks, highlighting the substantial potential of this paradigm in advancing autonomous driving and 3D scene understanding.

2. Related Works

3D Semantic Scene Completion. 3D Semantic Scene Completion (SSC) entails predicting occupancy and semantics for each voxel within a 3D scene, which was initially introduced by SSCNet [37]. Subsequent methods can be broadly categorized based on their model architectures and input modalities. Volume networks [19, 45] predominantly utilize Truncated Signed Distance Function (TSDF) features generated from depth data, and process them through 3D convolutional networks. On the other hand, view-volume networks [11, 20, 25, 33, 41] extract RGB or depth features before lifting them to 3D volumes. For a more in-depth overview of SSC, we refer readers to the survey by Roldão *et al*. [34].

Recently, camera-based SSC has garnered increasing attention for its immense applications in autonomous driving. MonoScene [3] presents the first purely visual solution, sampling RGB features along the line of sight and adapting a 3D UNet architecture. TPVFormer [16] introduces a Tri-Perspective View (TPV) representation to decompose voxels onto multiple view planes for efficient scene encoding. VoxFormer [22] proposes a two-stage framework that diffuses the global scene from proposed voxel features, resembling the Masked Autoencoder (MAE) [13]. OccFormer [46] applies a mask-wise prediction paradigm akin to MaskFormer [8, 9]. OccDepth [30] and NDC-Scene [44] improves 2D-to-3D geometric projection by leveraging stereo depth and Normalized Device Coordinates (NDC). OccNet [36] further envisions occupancy as a general scene descriptor for a wide scope of driving tasks.

In contrast to prior works, our proposed Symphonies differs by integrating instance queries to enhance scene model-

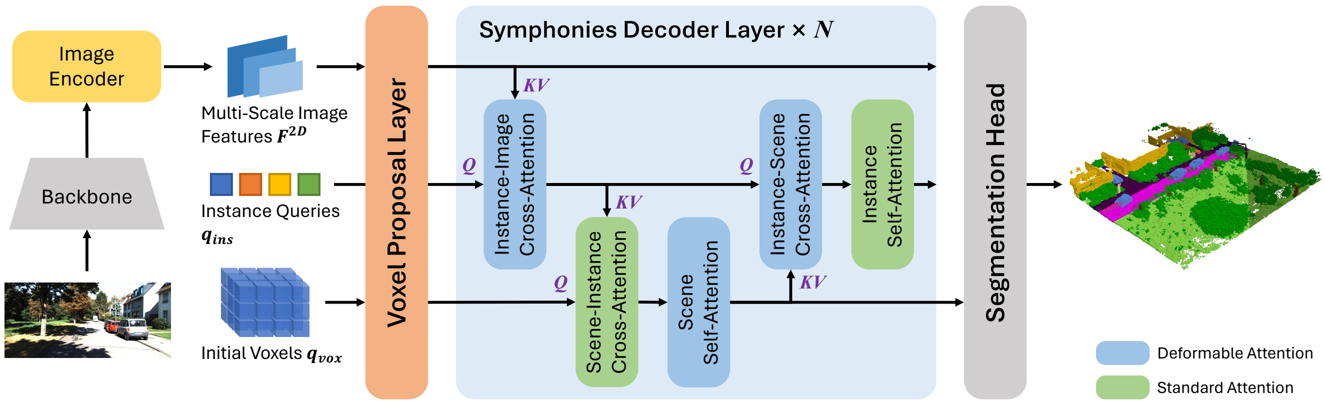

Figure 2. **Overview of Symphonies.** The Symphonies framework commences with extracting multi-scale image features through the image backbone and Instance-Aware Image Encoder. The Depth-Rectified Voxel Proposal Layer generates initial voxel features estimating the implicit surface. Subsequently, the Symphonies Decoder Layers, which consist of Serial Instance-Propagated Attentions, facilitate continuous interactions among the image, instances, and the scene, iterated N times. The Segmentation Head upsamples voxel features to the designated resolution and predicts class logits for each voxel.

ing with instance semantics and enriched contextual awareness, mitigating geometric ambiguities arising from voxel-wise modeling.

Camera-Based 3D Perception. The surge in autonomous driving applications has rekindled interest in camera-based 3D perception, owing to its cost-effectiveness and alignment with human visual perception. Early 3D object detection methods, such as FCOS3D [40] and DETR3D [42], straightforwardly extend 2D detectors to predict additional 3D bounding boxes. Among subsequent Transformer-based approaches, BEVFormer [23] and BEVDet [15] adopt the BEV space to align multi-frame features, while PolarDETR [7] establishes explicit correlations between image patterns. In addition, PETR [26] and PETRv2 [27] utilize 3D position embeddings to encode 2D features.

BEV segmentation, which is beneficial for representation learning and route planning, has also been explored. Approaches such as OFT [32], Lift-Splat [31], and FIERY [14] transform the camera plane into BEV via Inverse Perspective Mapping (IPM). PolarBEV [28] uses angle-specific and radius-specific embeddings to rasterize BEV features. BEVFormer [23] and CVT [47] aggregate BEV queries through cross-attention layers, while GKT [6] optimizes computational efficiency by constraining local attention calculations.

These developments are closely related to our work in SSC, where techniques like Deformable Attention [48] inspire our methodology to enhance 3D scene completion.

3. Scene from Instances

This section presents a comprehensive elaboration of our proposed Symphonies method, commencing with an architectural overview in Sec. 3.1. Subsequently, it proceeds to detail the Depth-Rectified Voxel Proposal Layer in Sec. 3.2 and the Symphonies Decoder Layer in Sec. 3.3, shedding light on their synergistic contributions. Further insights into training losses are discussed in Sec. 3.4.

3.1. Overview

The architectural details of our proposed Symphonies are illustrated in Fig. 2. In essence, Symphonies exclusively takes RGB images as input and extracts multi-scale 2D features F^{2D} through a ResNet-50 [12] image backbone and an Instance-Aware Deformable Transformer [48] Encoder, enhancing both global and instance semantics on the image plane. In the Symphonies Decoder, instance queries $q_{ins} \in \mathbb{R}^{N \times C}$ and the volumetric scene representation $q_{vox} \in \mathbb{R}^{C \times X \times Y \times Z}$ are initialized with learnable embeddings. Here, C signifies embedding dimensions, N denotes the number of instance queries, while X, Y, and Z indicate the scene grid dimensions.

The subsequent "scene-from-instances" process commences with the Depth-Rectified Voxel Proposal Layer initializing voxel proposals q_p with image features on the implicit surface. Multi-scale image features F^{2D}, scene features q_{vox}, and instance queries q_{ins} are then passed through our proposed Serial Instance-Propagated Attentions within the Symphonies Decoder Layers. This iterative process continuously propagates image features F^{2D} to scene features q_{vox} guided by instance queries q_{ins}, while simultaneously aggregating instance semantics from both modalities. The Segmentation Head then upsamples the scene features to the desired resolution, and predicts per-voxel class logits with a single linear layer after an Atrous Spatial Pyramid Pooling (ASPP) [5] module.

Depth Estimator. The depth prediction, obtained from a pre-trained depth estimator, is not explicitly illustrated in the diagram for clarity. It is employed to infer the implicit surface within the Voxel Proposal Layer and compute instance reference points in the scene volume. Specifically, we adopt the pre-trained Mobilestereonet [35] as the depth estimator, aligning with VoxFormer [22].

Instance-Aware Image Encoder. The Instance-Aware Image Encoder, vital for integrating instance semantics in the absence of direct instance-level supervision, employs a Deformable Transformer [48] adept at capturing long-range dependencies around diverse instances by attending to deformable reference points. Additionally, it is augmented by utilizing the pre-trained weight of MaskDINO [18] from panoptic segmentation [17], to enrich its instance awareness.

3.2. Depth-Rectified Voxel Proposal Layer

The Depth-Rectified Voxel Proposal Layer generates initial scene features for voxels located on the implicit surface, known as voxel proposals, which establishes coarse geometry awareness for subsequent instance-level aggregations. The implicit surface is computed through the conversion of camera coordinates to world coordinates using depth estimation, described as follows:

$$x^C = \mathcal{K}^{-1} \cdot (z_c \odot x^I) \quad (1)$$

$$x^W = [R, T]^{-1} \cdot x^C \quad (2)$$

where x^I, x^C, and x^W represent homogeneous coordinates of pixels, camera frustum, and the world, respectively. $\odot$ denotes the element-wise multiplication. The intrinsic matrix $\mathcal{K}$ encompasses camera parameters, while the extrinsic matrix is composed of the rotation matrix R and the translation vector T. z_c corresponds to the z-coordinate of the camera, *i.e.*, the depth estimation.

Based on the camera-to-world transformation, the positions V_p of voxel proposals are determined by mapping image points x^I to their corresponding world coordinates x^W, confined within the volume V:

$$V_p = \{x^W \mid x^W = \mathcal{T}^{IW}(x^I, z_c), \forall\, x^I \in I \text{ such that } x^W \in V\} \quad (3)$$

Here, $\mathcal{T}^{IW}$ refers to the camera-to-world transformation, I represents image pixels, and V represents voxel grids.

As illustrated in Fig. 3, the determined voxel features are initialized by aggregating multi-scale image features using Deformable Attention [48]. This process involves selecting the proposed voxels q_p associated with the positions V_p from scene volume q_{vox}, along with corresponding pixel positions p_I and 2D image features F^{2D}. This process is expressed as $q_p = \text{DeformAttn}(q_p, p_I, F^{2D})$.

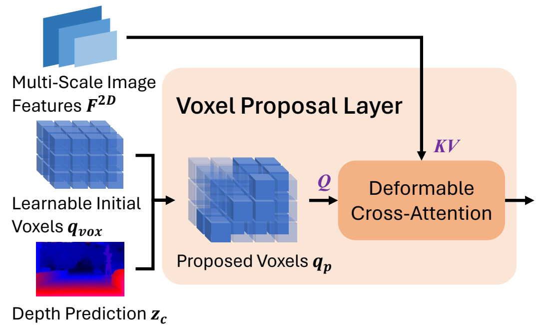

Figure 3. **Illustration of the Depth-Rectified Voxel Proposal Layer.**

The Deformable Attention operation, denoted as `DeformAttn`, dynamically aggregates query features q from features x with deformable reference points p_q. The mathematical expression is given by:

$$\text{DeformAttn}(q, p_q, x) = \sum_{k=1}^{K} A_{qk} W x(p_q + \Delta p_{qk}) \quad (4)$$

Here, K represents the number of sampling points, and A_{qk} stands for the learnable attention weight at sampling point k deformable based on queries q. The term Δp_{qk} denotes the offset applied to p_q, and W denotes the projection weight. The computation of multi-heads is omitted for brevity.

In contrast to the Query Proposal in VoxFormer [22], which employs an extra occupancy network [33] for generating coarse occupancy features, we refrain from it as it introduces additional geometric ambiguities in occlusion regions.

3.3. Symphonies Decoder Layer

The Symphonies Decoder Layer seamlessly integrates image features, instance queries, and voxel proposals, as depicted in Fig. 2. It orchestrates a dynamic flow of information, where instance queries serve as intermediaries to propagate extracted instance semantics to the broader scene representations. The process initiates with deformable cross-attention modules, which attend to the corresponding instance positions within image and scene features to extract instance-centric semantics. Subsequently, the instance self-attention and scene-instance cross-attention modules strengthen the internal cohesion of instances and aggregate scene context from instance queries. The scene self-attention mechanism further diffuses voxel features throughout the scene, especially for the occluded regions, as only visible surfaces are initially proposed.

The following paragraphs present a detailed explanation of the computations involved in their exact order of operation. To streamline the explanation, detailed discussions on certain components, including Feed-Forward Networks

(FFN), Layer Norms (LN), and identity connections, have been omitted.

Deformable Instance-Image Cross-Attention. For each instance query q_{ins}, deformable attention extracts surrounding features from multi-scale image features F^{2D} using learnable 2D reference points p^{2D}_{ins}, denoted as $q_{ins} = \text{DeformAttn}(q_{ins}, p^{2D}_{ins}, F^{2D})$.

Scene-Instance Cross-Attention. This attention mechanism aggregates scene features q_{vox} from instance queries, formulated as $q^{\in FOV}_{vox} = \text{CrossAttn}(q^{\in FOV}_{vox}, q_{ins}, q_{ins})$, where FOV refers to the "field of view" which is precomputed based on world-to-camera transformation excluding invisible voxels, reducing computational redundancy.

Deformable Scene Self-Attention. The scene self-attention enables feature propagation across the scene, where voxels attend to their neighbors: $q^{\in FOV}_{vox} = \text{DeformAttn}(q^{\in FOV}_{vox}, p_V, q_{vox})$. Here, p_V represents voxels' relative coordinates in the scene.

Deformable Instance-Scene Cross-Attention. Instance semantics are enhanced by integrating refined information from the reconstructed voxel features q_{vox}. Through coordinate transformation applied to 2D reference points, 3D reference points are derived as $p^{3D}_{ins} = \mathcal{T}^{IW}(p^{2D}_{ins})$. The instance-scene cross-attention is then formulated as $q_{ins} = \text{DeformAttn}(q_{ins}, p^{3D}_{ins}, q_{vox})$.

Instance Self-Attention. The instance self-attention captures internal relations and global context within instance queries, expressed as $q_{ins} = \text{SelfAttn}(q_{ins})$.

3.4. Losses

In the Symphonies framework, we adopt the Scene-Class Affinity Loss L_{scal} from MonoScene [3] to optimize precision, recall, and specificity concurrently. The Scene-Class Affinity Loss is applied to semantic and geometric predictions, in conjunction with the cross-entropy loss weighted by class frequencies. The overall loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}^{geo}_{scal} + \mathcal{L}^{sem}_{scal} + \mathcal{L}_{ce} \tag{5}$$

Following the DETR series [4], auxiliary losses are applied after each Symphonies Decoder Layer for enhanced supervision, following the same formulation as $\mathcal{L}$ but scaled by a factor of 0.5.

4. Experiments

In this section, we present the evaluation results of our proposed Symphonies on SemanticKITTI [1] and SSCBench-KITTI-360 [21] datasets. The comparative analysis positioning Symphonies against existing approaches is detailed in Sec. 4.3. Additionally, comprehensive ablation studies are conducted in Sec. 4.4 to shed light on the thorough understanding of Symphonies.

4.1. Dataset and Metric

The evaluation is performed on SemanticKITTI [1] and SSCBench-KITTI-360 [21] datasets, both providing densely annotated urban driving scene sequences, 22 and 9 respectively, from the KITTI Odometry Benchmark [10]. These datasets voxelize point clouds and label scenes measuring $51.2m \times 51.2m \times 64m$, with voxel grids of $256 \times 256 \times 32$ and a voxel size of 0.2m. SemanticKITTI comprises 10 sequences for training, 1 sequence for validation, and 11 sequences for testing. It furnishes RGB images with shapes of 1226×370 as inputs and encompasses 20 semantic classes. SSCBench-KITTI-360 provides 7 sequences for training, 1 sequence for validation, and 1 sequence for testing, with 19 semantic classes and RGB images of size 1408×376. For our camera-based approach, we exclusively adopt RGB images as input, and report the intersection over union (IoU) and mean IoU (mIoU) metrics aligned with standard practices. The IoU metric assesses the binary classification of empty versus occupied voxels, reflecting the performance of geometric reconstruction. Conversely, the mIoU metric provides a comprehensive assessment of semantic understanding, making it the primary metric in most benchmarks.

4.2. Implementation Details

In line with prior studies [3, 16, 22], we train Symphonies for 30 epochs on 4 NVIDIA 3090 GPUs, with a batch size of 4 images. We apply random horizontal flip augmentation and employ the AdamW [29] optimizer with an initial learning rate of 2e-4 and a weight decay of 1e-4. Learning rate reduction occurs by a factor of 0.1 at the 25th epoch. The ResNet-50 [12] backbone and Image Encoder are initialized with pre-trained MaskDINO [18] weights.

4.3. Main Results

We conduct a comprehensive comparison of Symphonies with the latest state-of-the-art camera-based methodologies on the SemanticKITTI and SSCBench-KITTI-360 datasets. The results, outlined in Tab. 1 and Tab. 2, establish the superior performance of Symphonies with substantial improvements of 2.72 and 4.77 mIoU on SemanticKITTI and SSCBench-KITTI-360, respectively. Specifically, Symphonies showcases particular excellence in instance classes,

Method	IoU	mIoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd. (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)
LMSCNet† [33]	31.38	7.07	46.70	19.50	13.50	3.10	10.30	14.30	0.30	0.00	0.00	0.00	10.80	0.00	10.40	0.00	0.00	0.00	5.40	0.00	0.00
AICNet† [20]	23.93	7.09	39.30	18.30	19.80	1.60	9.60	15.30	0.70	0.00	0.00	0.00	9.60	1.90	13.50	0.00	0.00	0.00	5.00	0.10	0.00
JS3C-Net† [43]	34.00	8.97	47.30	21.70	19.90	2.80	12.70	20.10	0.80	0.00	0.00	4.10	14.20	3.10	12.40	0.00	0.20	0.20	8.70	1.90	0.30
MonoScene* [3]	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [16]	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	**3.70**	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
VoxFormer [22]	**42.95**	12.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	**2.60**	0.20	11.10	5.10	4.90
OccFormer [46]	34.53	12.32	55.90	**30.30**	**31.50**	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
Symphonies	42.19	**15.04**	**58.40**	29.30	26.90	**11.70**	**24.70**	**23.60**	3.20	**3.60**	**2.60**	**5.60**	**24.20**	**10.00**	**23.10**	**3.20**	1.90	**2.00**	**16.10**	**7.70**	**8.00**

Table 1. **Quantitative results on SemanticKITTI `test`.** † denotes the results provided by [3]. * represents the reproduced results in [16, 46]. The best results are in **bold**.

Method	IoU	Prec.	Rec.	mIoU	car (2.85%)	bicycle (0.01%)	motorcycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	parking (2.31%)	sidewalk (6.43%)	other-grnd. (2.05%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrain (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-struct. (4.33%)	other-obj. (0.28%)
LiDAR-based methods																						
SSCNet [37]	**53.58**	69.63	**69.92**	16.95	**31.95**	0.00	0.17	10.29	0.00	0.07	**65.70**	**17.33**	**41.24**	3.22	**44.41**	6.77	**43.72**	**28.87**	0.78	0.75	8.69	0.67
LMSCNet [33]	47.53	**72.77**	57.55	13.65	20.91	0.00	0.00	0.26	0.58	0.00	62.95	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
Camera-based methods																						
MonoScene [3]	37.87	56.73	53.26	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [16]	40.22	59.32	<u>55.54</u>	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
VoxFormer [22]	38.76	58.52	53.44	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.18	4.97	28.99	14.69	6.51	6.92	3.79	2.43
OccFormer [46]	40.27	59.70	55.31	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	<u>36.42</u>	4.80	31.00	<u>19.51</u>	7.77	8.51	6.95	4.60
Symphonies	<u>44.12</u>	<u>69.24</u>	54.88	**18.58**	<u>30.02</u>	**1.85**	**5.90**	**25.07**	**12.06**	**8.20**	<u>54.94</u>	<u>13.83</u>	<u>32.76</u>	**6.93**	35.11	**8.58**	<u>38.33</u>	11.52	**14.01**	**9.57**	**14.44**	**11.28**

Table 2. **Quantitative results on SSCBench-KITTI-360 `test`.** The results for counterparts are provided in [21]. The best results among all methods are in **bold**, and the best results for camera-based methods are <u>underlined</u>.

e.g., buildings, cars, *etc.* This underscores its prowess in capturing and modeling intricate instance semantics. While VoxFormer attains a marginally higher IoU on SemanticKITTI, its adoption of two-stage training and extra occupancy prediction network disrupts end-to-end training and introduces additional geometric ambiguities. This complexity hampers its robustness, especially on KITTI-360.

The superiority of Symphonies becomes more pronounced on SSCBench-KITTI-360, which can be attributed to the ample data samples and high-quality annotations. Moreover, Symphonies even outperforms LiDAR-based methods in terms of mIoU, despite LiDAR's inherent advantage in IoU owing to its more precise position awareness, particularly at a distance.

4.4. Ablation Studies

The ablation analysis is conducted on the SemanticKITTI validation set from four key perspectives: overall architectural components, the Symphonies Decoder, the Voxel Proposal Layer, and the Image Encoder.

Ablation on architectural components. Tab. 3 presents the breakdown analysis of various architectural components within Symphonies. The vanilla baseline can be considered as a light-weight alternative to MonoScene, composed of a ResNet-50 backbone, an Image Encoder without pretrained weight, a 2D-to-3D projection via FLoSP [3], and a single 3D ASPP layer as the 3D decoder, Pre-training the Image Encoder leads to a notable improvement of 2.15 mIoU, emphasizing the effectiveness of instance awareness brought by 2D segmentation pre-training. Further, the proposed Depth-Rectified Voxel Proposal Layer improves performance by 0.75 mIoU through more accurate geometry. The Symphonies Decoder significantly boosts performance by 5.38 IoU, attributed to its dynamic instance modeling and context-capturing capabilities. In summary, the analy-

Method	IoU	mIoU	Params (M)	FLOPs (G)
Baseline	34.06	10.44	57.22	529.20
+ Pre-trained Encoder	35.97 (+1.91)	12.59 (+2.15)	57.22	529.20
+ Voxel Proposal Layer	36.54 (+0.57)	13.34 (+0.75)	57.42	535.84
+ Symphonies Decoder	**41.92** (+5.38)	**14.89** (+1.55)	59.31	611.89

Table 3. **Ablation study on architectural components in Symphonies.**

sis in Tab. 3 affirms the effectiveness of the proposed components in Symphonies.

Ablation on the Symphonies Decoder. To gain insights into the functionality of contextual instance queries, we assess the modular interactions within the Symphonies Decoder Layer. As depicted in Tab. 4, the incorporation of instance queries with either instance-image or instance-scene cross-attention considerably enhances performance by over 5 IoU and about 1 mIoU. This substantiates the significance of instance queries for adaptive aggregation of instance semantics. Among them, the instance-image cross-attention brings less improvement, suggesting that original image features have already been adequately captured in the preceding Voxel Proposal Layer. The instance self-attention further improve the performance, underlining the contextual effectiveness of efficient fusion among instance queries. Besides, though Scene SA may seem to only marginally improve the performance, its omission reveals significant fluctuations during training, attributed to the sparse interactions within instance deformable attentions. This underscores the pivotal role of Scene SA in fostering the generation of consistent features.

Scn.-Inst. CA	Inst.-Img. CA	Inst.-Scn. CA	Inst. SA	Scn. SA	IoU	mIoU
					35.97	13.34
✓	✓				41.36	14.02
✓		✓			41.44	14.37
✓		✓	✓		41.35	14.63
✓	✓	✓	✓		41.75	14.73
✓	✓	✓	✓	✓	**41.92**	**14.89**

Table 4. **Ablation study on Symphonies Decoder. SA: Self-Attention, CA: Cross-Attention.**

Ablation on the Voxel Proposal Layer. Comparing the Depth-Rectified Voxel Proposal Layer (VPL) with FLoSP from MonoScene [3] casting pixels to voxels along the line of sight, as well as the mono VPL using monocular depth estimator AdaBins [2] (0.058 REL on KITTI), we note significant occupancy prediction improvements using the stereo VPL based on MobileStereoNet [35] (0.66 EPE on KITTI 2015), as shown in Tab. 5. This indicates that the rectification of more precise depth estimation contributes to mitigating geometric ambiguities, aligning with the findings in VoxFormer.

2D-to-3D Projection	IoU	mIoU
FLoSP [3]	36.02	11.96
VPL (mono)	38.37	12.20
VPL (stereo)	**41.92**	**14.89**

Table 5. **Ablation on the Depth-Rectified Voxel Proposal Layer.**

Ablation on the Instance-Aware Image Encoder. Tab. 6 evidently showcases the synergistic effects of utilizing pre-trained weights for the Image Encoder with instance queries. Solely utilizing pre-trained weights from MaskDINO [18] contributes an additional improvement of 0.48 mIoU. Moreover, incorporating instance queries with the pre-trained encoder yields a significant improvement of 1.36 mIoU, implying that the proposed instance queries benefit from the enhanced instance awareness of the encoder.

Pre-trained Encoder	Inst. Queries	IoU	mIoU
		41.09	13.53
	✓	41.42	13.32
✓		41.18	14.01
✓	✓	**41.92**	**14.89**

Table 6. **Ablation study on the Instance-Aware Image Encoder.**

4.5. Visualizations

Qualitative Results. Fig. 4 presents the visualizations of Symphonies on SemanticKITTI `val`, in comparison to the counterpart MonoScene. Symphonies generates more detailed predictions for instance-centric classes such as cars and trunks, as well as preserves clear and coherent layouts for structures like buildings and vegetation, attributed to the enriched instance semantics and contextual information provided by instance queries. In contrast, MonoScene

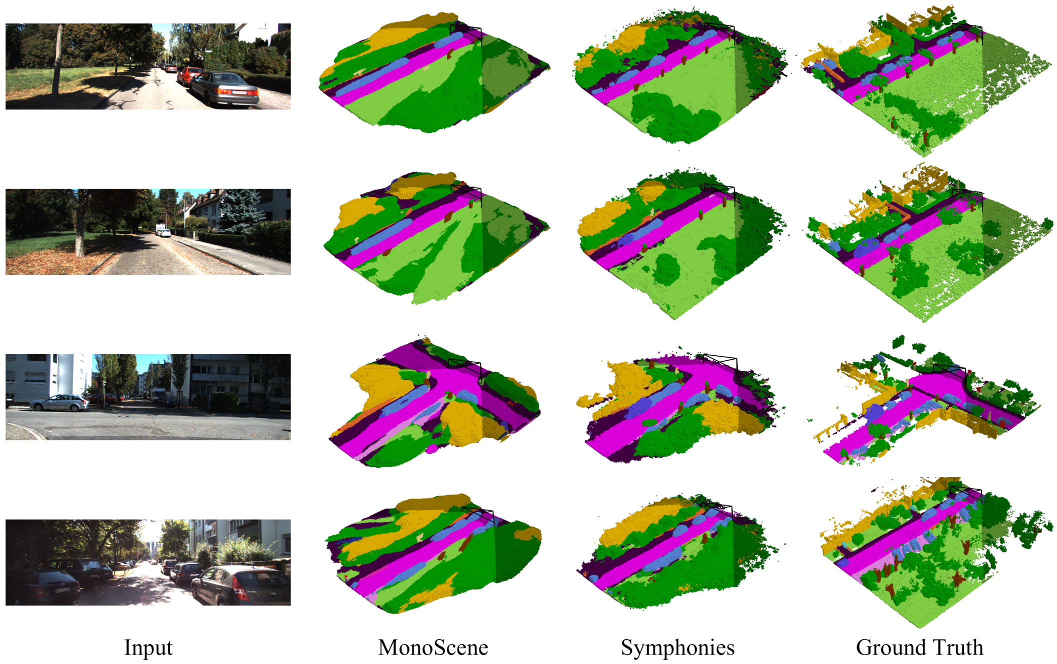

Figure 4. **Qualitative visualizations on SemanticKITTI `val`.** Symphonies consistently produces detailed predictions for objects such as cars and trunks, while maintaining coherent layouts for structures like buildings and vegetation.

produces vague predictions with a radial shape, which is indicative of the aforementioned ambiguous geometry. These results underscore the superior capability of Symphonies in capturing fine-grained scene representations and enhancing overall scene understanding.

Attention Map Analysis. The attention map analysis in Fig. 5 provides insights into the mechanisms of the Serial Instance-Propagated Attentions within Symphonies layers. Notably, instance queries exhibit selective attention to corresponding regions in both the image and the scene. Additionally, they activate the semantically related regions within the scene-instance cross-attention. This observation validates the effect of our claimed instance-centric semantics in facilitating effective scene modeling.

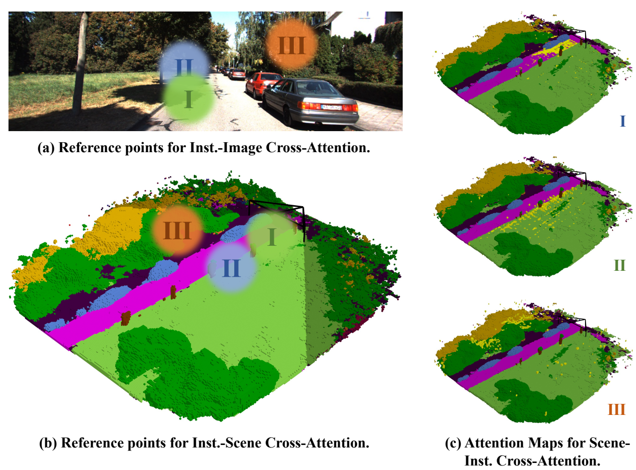

Figure 5. **Analysis of attention maps within Symphonies.**

5. Conclusion

In this paper, we introduced Symphonies, a novel paradigm for 3D Semantic Scene Completion. Symphonies effectively integrates instance-centric semantics and scene context from both images and volumes, addressing the limitations posed by geometric ambiguity in prior voxel-wise modeling methods. Extensive experiments demonstrate the superiority of our approach over existing methods. We anticipate Symphonies to inspire future research and contribute to advancements in autonomous driving and 3D perception.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62276108.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Int. Conf. Comput. Vis.*, pages 9296–9306,

2019. 2, 5

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7

[3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3981–3991, 2022. 1, 2, 5, 6, 7

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020. 5

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 834–848, 2018. 3

[6] Shaoyu Chen, Tianheng Cheng, Xinggang Wang, Wenming Meng, Qian Zhang, and Wenyu Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*, 2022. 1, 3

[7] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022. 3

[8] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, pages 17864–17875, 2021. 2

[9] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1280–1289, 2022. 2

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012. 5

[11] Yuxiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *IJCAI*, pages 726–732, 2018. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 3, 5

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15979–15988, 2022. 2

[14] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: future instance prediction in bird's-eye view from surround monocular cameras. In *Int. Conf. Comput. Vis.*, pages 15253–15262, 2021. 3

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

[16] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9223–9232, 2023. 1, 2, 5, 6

[17] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 4

[18] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: towards A unified transformer-based framework for object detection and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3041–3050, 2023. 4, 5, 7

[19] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian D. Reid. RGBD based dimensional decomposition residual network for 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7693–7702, 2019. 2

[20] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3348–3356, 2020. 2, 6

[21] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023. 2, 5, 6

[22] Yiming Li, Zhiding Yu, Christopher B. Choy, Chaowei Xiao, José M. Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9087–9098, 2023. 2, 4, 5, 6

[23] 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 *Eur. Conf. Comput. Vis.*, pages 1–18, 2022. 1, 3

[24] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 3292–3310, 2023. 2

[25] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *Adv. Neural Inform. Process. Syst.*, pages 261–272, 2018. 2

[26] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: position embedding transformation for multi-view 3d object detection. In *Eur. Conf. Comput. Vis.*, pages 531–548, 2022. 3

[27] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 3

[28] Zhi Liu, Shaoyu Chen, Xiaojie Guo, Xinggang Wang, Tianheng Cheng, Hongmei Zhu, Qian Zhang, Wenyu Liu, and Yi Zhang. Vision-based uneven BEV representation learning with polar rasterization and surface estimation. In *Conference on Robot Learning*, pages 437–446, 2022. 3

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019. 5

[30] 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, 2

[31] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Eur. Conf. Comput. Vis.*, pages 194–210, 2020. 1, 3

[32] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *Brit. Mach. Vis. Conf.*, page 285, 2019. 1, 3

[33] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, pages 111–119, 2020. 2, 4, 6

[34] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *Int. J. Comput. Vis.*, 130(8):1978–2005, 2022. 2

[35] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *WACV*, pages 677–686, 2022. 4, 7

[36] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. *arXiv preprint arXiv:2306.02851*, 2023. 2

[37] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 190–198, 2017. 2, 6

[38] 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

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 1

[40] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: fully convolutional one-stage monocular 3d object detection. In *ICCVW*, pages 913–922, 2021. 3

[41] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *Int. Conf. Comput. Vis.*, pages 8607–8616, 2019. 2

[42] Yue Wang, Vitor 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, 2021. 3

[43] 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 *AAAI*, pages 3101–3109, 2021. 6

[44] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. *arXiv preprint arXiv:2309.14616*, 2023. 2

[45] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Int. Conf. Comput. Vis.*, pages 7800–7809, 2019. 2

[46] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 1, 2, 6

[47] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13750–13759, 2022. 3

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *Int. Conf. Learn. Represent.*, 2021. 1, 3, 4