

Disentangling Instance and Scene Contexts for 3D Semantic Scene Completion

Enyu Liu* En Yu* Sijia Chen Wenbing Tao[†]
Huazhong University of Science and Technology
{eyliu, yuen, sijiachen, wenbingtao}@hust.edu.cn

Abstract

3D Semantic Scene Completion (SSC) has gained increasing attention due to its pivotal role in 3D perception. Recent advancements have primarily focused on refining voxel-level features to construct 3D scenes. However, treating voxels as the basic interaction units inherently limits the utilization of class-level information, which is proven critical for enhancing the granularity of completion results. To address this, we propose *Disentangling Instance and Scene Contexts (DISC)*, a novel dual-stream paradigm that enhances learning for both instance and scene categories through separated optimization. Specifically, we replace voxel queries with discriminative class queries, which incorporate class-specific geometric and semantic priors. Additionally, we exploit the intrinsic properties of classes to design specialized decoding modules, facilitating targeted interactions and efficient class-level information flow. Experimental results demonstrate that DISC achieves state-of-the-art (SOTA) performance on both SemanticKITTI and SSCBench-KITTI-360 benchmarks, with mIoU scores of 17.35 and 20.55, respectively. Remarkably, DISC even outperforms multi-frame SOTA methods using only single-frame input and significantly improves instance category performance, surpassing both single-frame and multi-frame SOTA instance mIoU by 17.9% and 11.9%, respectively, on the SemanticKITTI hidden test. The code is available at <https://github.com/Enyu-Liu/DISC>.

1. Introduction

With the rapid development of autonomous driving [11, 37], accurate environmental perception is crucial for tasks like navigation and obstacle avoidance [15]. The 3D Semantic Scene Completion (SSC) task addresses these critical demands by jointly predicting scene geometry and semantics to enable comprehensive 3D scene understanding. SSC methods are primarily categorized into lidar-based [6, 31, 45] and vision-based approaches [2, 12, 13, 25, 28, 50].

*Equal contribution.

[†]Corresponding author.

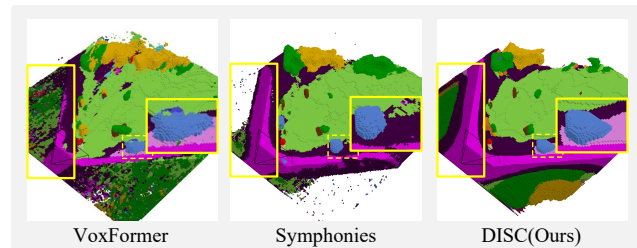


Figure 1. **Comparative Analysis of Different Methods.** Incorporating instance or scene specific information significantly improves prediction accuracy for their corresponding categories.

Owing to their lower memory consumption, vision-based methods have gained prominence. VoxFormer [25] pioneers a sparse-to-dense architecture that mitigates projection blurring, and subsequent advances enhance voxel feature learning through techniques like self-distillation [38], context priors [50], and implicit instance fusion [13]. Since all these methods utilize voxels as the basic units for feature interaction, we classify them as voxel-based methods.

Although voxel-based methods have achieved remarkable performance, we observe distinct prediction flaws in instance and scene categories: instance categories suffer from class omission and semantic ambiguity due to occlusion and projection errors, while scene categories exhibit structural incoherence from out-of-view regions. Based on this analysis, we argue that fully leveraging class-level information can help address the divergent challenges of different categories. As shown in Fig. 1, compared to VoxFormer [25] (no category-specific priors) and Symphonies [13] (implicit instance priors), our method explicitly integrates both instance and scene information, achieving superior performance in completion details. Nevertheless, despite the proven benefits of class-level information, voxel-based works have not further extended the class-level concepts. We attribute this limitation to three critical challenges faced by voxel-based methods in leveraging class-level information: (1) Voxel construction process inherently disrupts the structural information of categories; (2) Voxel-based methods require a unified module to handle features across all categories, limiting the ability to address divergent challenges posed by instance and scene categories; (3)

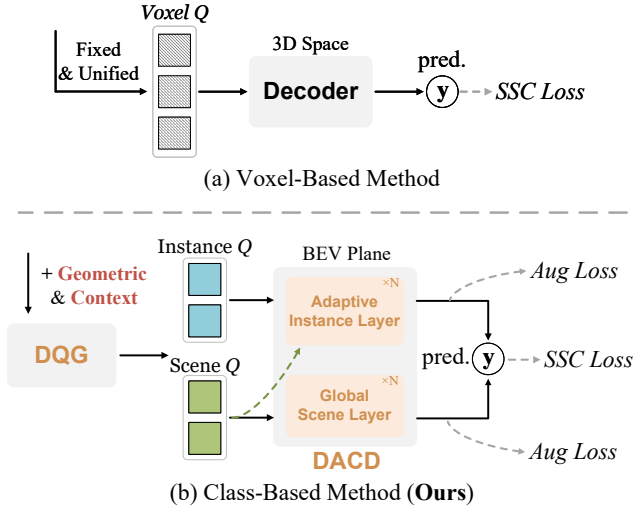


Figure 2. **Comparison of main architectures.** The key differences between our Class-Based Method and the previous Voxel-Based Method are highlighted in red-brown. Our method initializes instance and scene queries with semantic and geometric priors, using a dual-stream structure and tailored modules for class-discriminative scene semantic completion.

Voxel-based methods rely on 3D spatial feature interactions, which impose high computational costs and hinder further integration of class-level information.

To overcome these challenges, we depart from conventional voxel-based paradigms and propose **Disentangling Instance and Scene Contexts (DISC)**, a class-aware dual-stream architecture for discriminative reconstruction. As shown in Fig. 2(b), DISC operates in BEV space and introduces two core modules to leverage category-specific information: a Discriminative Query Generator (DQG) and a Dual-Attention Class Decoder (DACD). Our method addresses three critical issues: (1) Replaces voxel queries with independent instance and scene queries, preserving category-specific structural information while integrating richer geometric and semantic priors; (2) Employs tailored modules to resolve distinct instance and scene challenges; (3) Mitigates BEV’s height-axis limitations by disentangling instance and scene contexts, thereby enabling efficient class-level integration with lower computational costs.

Specifically, our DQG module first initializes independent queries with positional priors based on instance and scene properties, followed by aggregating contextual information to provide semantic priors. Moreover, we argue that feature disentanglement mitigates the BEV limitations in height-axis modeling. For example, road and pedestrian categories within the same BEV grid exhibit significant height disparities, causing vertical feature ambiguity during 3D reconstruction. However, disentangling instance and scene contexts alleviates such multi-height distributions in shared grids. Leveraging this insight, DISC adopts BEV

space for query initialization and feature interaction, which significantly reduces memory usage. To address projection errors, we further propose a Coarse-to-fine BEV generation module to deliver precise semantic priors.

Moreover, to design targeted solutions, we further analyze the root causes of instance and scene errors and their relationships. We identify that scene categories suffer from topological errors due to weak global reasoning (e.g., roads appearing in terrain), while instance categories exhibit semantic ambiguities from projection errors and occlusion-induced feature loss. Additionally, scene context aids instance reasoning, for example, pole-shaped objects on sidewalks are more likely to be traffic lights than trunks. Based on these analysis, our DACD decoder implements two specialized layers: the Adaptive Instance Layer (AIL) dynamically fuses image features and scene context to mitigate projection errors and recover occluded details, and the Global Scene Layer (GSL) establishes cross-region interactions to expand the scene’s receptive field and ensure layout coherence. Both layers leverage class-specific properties to regulate feature propagation within their respective streams.

Finally, we evaluate DISC on SemanticKITTI and SSCBench-KITTI-360 benchmarks, achieving state-of-the-art (SOTA) mIoU scores of 17.35 and 20.55 respectively. Notably, with only single-frame inputs, DISC outperforms multi-frame SOTA methods. Furthermore, DISC demonstrates substantial gains in instance-level understanding, improving instance mIoU by **17.9%** over SOTA methods, while simultaneously achieving SOTA scene mIoU performance on the SemanticKITTI hidden test set.

In summary, our main contributions are the following:

- We propose **DISC**, a novel class-based architecture for the SSC task, featuring a dual-stream structure that decouples instance and scene categories in the BEV plane to fully utilize class-level information.
- We propose the Discriminative Query Generator (**DQG**) that provides geometric and semantic priors for both instance and scene queries, paired with the Dual-Attention Class Decoder (**DACD**) to address the distinct challenges of instance and scene categories, ensuring accurate and efficient feature interaction within each stream.
- We achieve SOTA performance with DISC on both SemanticKITTI and SSCBench-KITTI-360 benchmarks. Moreover, DISC is the first to outperform multi-frame SOTA methods using only single-frame inputs.

2. Related Works

3D Semantic Scene Completion. The 3D Semantic Scene Completion (SSC) task, introduced by SSCNet [36], is an important task in computer vision [4, 5, 21, 47, 48] that aims to jointly estimate the complete geometry and semantics of a scene from sparse inputs, playing a crucial role in applications like autonomous driving and virtual reality.

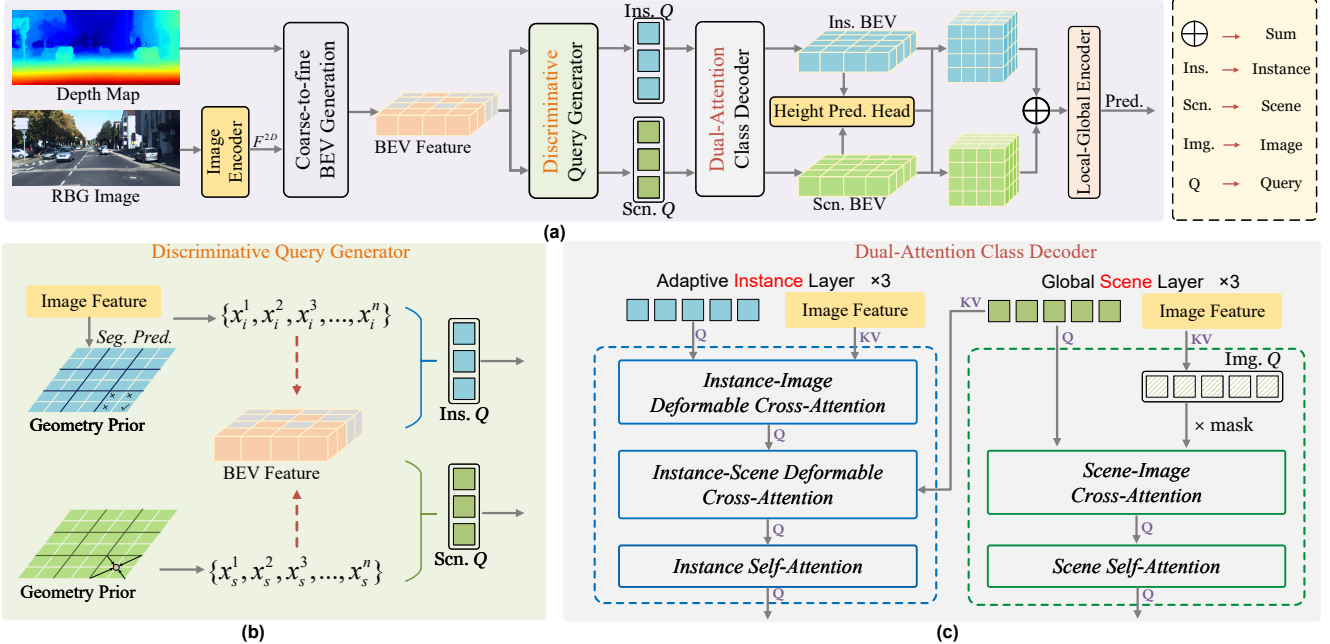


Figure 3. **The overall architecture.** (a) DISC is a novel semantic scene completion method with a dual-stream framework for specialized instance and scene categories processing. (b) The Discriminative Query Generator (DQI) integrates geometric and contextual priors into instance and scene queries based on category attributes. (c) Details of the Adaptive Instance Layer (AIL) and the Global Scene Layer (GSL), which address the distinct challenges faced by instance and scene categories during the reconstruction process in a differentiated manner. For clarity, the Feed-Forward Network (FFN) and positional embedding are omitted in the figure.

Early methods [6, 31, 45] focused on LiDAR for 3D semantic occupancy prediction, but camera-based approaches have gained prominence due to their cost efficiency. Recent voxel-based methods, such as MonoScene [2], TPVFormer [12], and VoxFormer [25], enhance semantic information for each voxel, using techniques like monocular image inputs, tri-perspective views, and deformable attention mechanisms. Other approaches like OccDepth [28] and VPOcc [16] improve geometric projection and address perspective distortion, while Symphonize [13] and CGFormer [50] focus on instance-level representations and enhanced voxel features. Voxel-based methods struggle to distinguish instance and scene features, while our approach employs class-aware modeling to resolve this divergence.

BEV-based 3D Perception BEV-based methods have gained significant attention in 3D perception due to their compact scene representation [20]. These methods can be classified into transformer-based and frustum-based approaches, depending on how BEV maps are derived from camera images. Transformer-based methods [3, 7, 8, 14, 26, 44] project BEV grids onto camera images and pull features from neighboring points into the BEV space. Frustum-based methods [22, 23, 29, 30, 40, 51] lift image features into 3D frustums by predicting depth probabilities and then use voxel pooling to generate BEV features. In this work, we build upon the Lift-Splat-Shoot (LSS) method [29] and leverage depth information to reduce geometric ambiguity

caused by depth estimation errors. While BEV cuts computational costs [10, 49] for 3D scene completion, it underperforms scene-based methods due to vertical coupling of instance and scene features in unified BEV representations. Our feature-decoupling modules resolve this height-dimension constraint, unlocking BEV’s potential.

3. Method

3.1. Overall Architecture

The overall architecture of DISC is shown in Fig. 3. A 2D backbone [13] extracts multi-scale feature maps F^{2D} , followed by the Discriminative Query Generator (DQG) producing instance queries $\mathbf{Q}_{\text{ins}} \in \mathbb{R}^{N_{\text{ins}} \times C}$ and scene queries $\mathbf{Q}_{\text{scn}} \in \mathbb{R}^{N_{\text{scn}} \times C}$ in the BEV space, where N_{ins} and N_{scn} represent the number of initial queries (Sec. 3.2). These queries are processed by the Dual-Attention Class Decoder (DACD) for class-specific feature interaction (Sec. 3.3). Finally, the refined instance and scene features are fused in 3D space and passed through a prediction head to obtain the scene completion results (Sec. 3.4). Detailed descriptions of each component are provided in the following subsections.

Depth Estimator. Following previous works [13, 19, 50], we employ a pretrained MobileStereoNet [35] to estimate the depth of the input image. The estimated depths are used to supervise depth predictions and compute the 3D positions

of projected points from pixel coordinates.

3.2. Discriminative Query Generator

As shown in Fig. 3(b), we propose a category-aware query design that distinguishes between instance queries and scene queries to preserve their distinct geometric properties. Unlike uniform voxel queries, our approach integrates geometric and semantic priors into query initialization while maintaining computational efficiency. First, BEV features are derived by projecting 2D features (F^{2D}) onto the BEV plane, with projection errors mitigated through a Coarse-to-fine BEV Generation module. For instances, their sparse spatial distribution and projection errors make the image space more suitable for identifying potential positions. However, perspective distortion complicates uniform sampling across different distances. Therefore, we first locate potential instances in the image space and then refine them in the BEV space. For scene queries, we adopt a large-sized design to capture continuous spatial distributions while preserving structural coherence, thereby avoiding the segmentation artifacts caused by fine-grained voxel queries. Notably, all queries incorporate BEV positional embeddings (detailed in the appendix).

Coarse-to-fine BEV Generation. To address projection errors in LSS [29], we first generate coarse voxel features $V_{\text{coarse}} \in \mathbb{R}^{C \times X \times Y \times Z}$ via lifting, where X , Y , and Z correspond to the scene grid dimensions. Depth-guided surface voxels are then refined using F^{2D} , producing optimized $V_{\text{fine}} \in \mathbb{R}^{C \times X \times Y \times Z}$ through feature fusion. Finally, BEV features $C \in \mathbb{R}^{C \times X \times Y}$ are obtained by Z-axis max pooling of V_{fine} . Further implementation details are provided in the supplementary materials.

Instance Query. To improve the recall of instance categories, we detect potential instances through image-space segmentation and project candidate pixels to BEV. A parallelizable neighbor suppression strategy selects N_{ins} reference positions: $k \times k$ grids form voting blocks, with highest-probability candidates retained per block. This emphasizes small and long-tail objects while maintaining efficiency, expressed as:

$$X_{\text{ins}} = \{\text{CT}(\mathbf{g}_n) \mid \mathbf{g}_n \in \text{Top-}N(\{\text{Max}(B_{k \times k}^i)\}_{i=1}^s)\} \quad (1)$$

Here, s denotes the number of blocks, and $B_{k \times k}^i$ represents the i -th block of $k \times k$ grids, where i ranges from 1 to s . The Max operation selects the grid with the highest probability in each block as a candidate. \mathbf{g}_n refers to the top- N selected candidate grids, and $\text{CT}(\mathbf{g}_n)$ denotes their center coordinates, forming the reference set X_{ins} for instance queries. The features at these locations in BEV plane initialize the instance queries: $\mathbf{Q}_{\text{ins}} = C[X_{\text{ins}}]$.

Scene Query. For scene categories, we design a patch-based scene query generation strategy to enrich the informa-

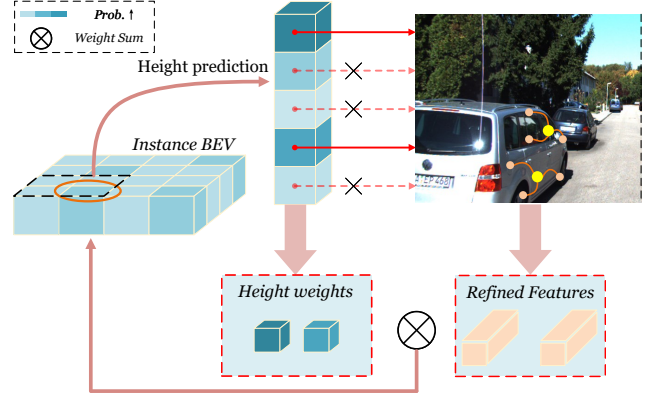


Figure 4. **Instance-Image Cross-Attention.** For each instance query, we adaptively select a series of heights and combine them with its reference point coordinates on the BEV plane to project the query into the image space. This enables capturing image features across multiple height levels.

tion in each individual query. Specifically, BEV features are divided into multiple equally sized patches, with the center point of each patch serving as the reference point for a scene query. Each patch is upsampled via convolution until the feature size reaches 1×1 , which is then used as the initialization feature for the scene query, expressed as:

$$\mathbf{Q}_{\text{scn}} = \text{UpSample}(C) \quad (2)$$

Here, UpSample represents the upsampling operation.

3.3. Dual-Attention Class Decoder

As discussed in Sec. 1, existing methods [13, 25, 50] employ a unified attention module for both instance and scene features, limiting targeted optimizations. Given the distinct characteristics between instances and scenes, this approach is suboptimal. In contrast, we design task-specific decoding layers tailored to each challenges, as shown in Fig. 3 (c). Prone to occlusion and projection errors, instances require multi-level geometric and semantic information for robust reasoning. For example, height direction features and contextual categories facilitate instance refinement and ambiguity reduction. Scene categories, with their continuous distribution, require global receptive fields to capture spatial relationships and maintain layout coherence. Based on these insights, we propose the Dual-Attention Class Decoder comprising an Adaptive Instance Layer and a Global Scene Layer. As follows, we will explain how they achieve targeted feature propagation and interaction.

Adaptive Instance Layer. As shown in Fig. 4, to address the loss of height information in BEV space, the instance query is transformed into pillar-like queries. Given the significant height variance within instance classes (e.g., trucks and traffic signs are much taller than pedestrians), we use adaptive height sampling rather than uniform sampling. For

each instance query q_{ins} , discrete candidate heights are initialized. A linear layer predicts probabilities for these candidates, and TOP-N selection identifies the most probable heights. These heights are combined with the reference coordinates x_{ins} to form reference points $P_j = (x_{\text{ins}}, h_j)$. These points are then projected onto the image feature space, where nearby features are sampled using deformable cross attention. The weighted sum of these features is then used to update q_{ins} , expressed as :

$$q_{\text{ins}} = \sum_{j=1}^N w_j \text{DA}(q_{\text{ins}}, F^{2D}, \mathcal{T}^{WI}(x_{\text{ins}}, h_j)) \quad (3)$$

Here, h_j represents a sampled height, with a total of N heights, and w_j is the corresponding sampling weight. \mathcal{T}^{WI} denotes the projection from the world coordinate system to image space for feature sampling. In practice, the 3D reference points are randomly offset within the voxel grid range to enhance the ability of q_{ins} to capture features.

After that, the instance queries extract scene features from the region of interest through attention mechanism, which is formulated as:

$$q_{\text{ins}} = \text{DA}(q_{\text{ins}}, C_{\text{scn}}, x_{\text{ins}}) \quad (4)$$

where C_{scn} represents scene features output by the GSL. We also utilize self attention to capture internal relationships among instance queries and recover lost information:

$$\mathbf{Q}_{\text{ins}} = \text{SelfAttn}(\mathbf{Q}_{\text{ins}}) \quad (5)$$

Finally, instance features are propagated across the entire BEV plane using a UNet-like [32] network, yielding C_{ins} .

Global Scene Layer. To enable each scene query to capture global information, we construct \mathbf{Q}_{img} following the same method as \mathbf{Q}_{scn} , obtained by the smallest scale image feature F_s^{2D} . For each scene query q_{scn} , global semantic features are aggregated from \mathbf{Q}_{img} via cross attention, as shown below:

$$q_{\text{scn}} = \text{CrossAttn}(q_{\text{scn}}, \mathbf{Q}_{\text{img}}^{\in \text{Mask}}, \mathbf{Q}_{\text{img}}^{\in \text{Mask}}) \quad (6)$$

where Mask is a random mask that discards a portion of \mathbf{Q}_{img} , simulating information loss (e.g., occlusions), which aids the network in better inferring the scene layout.

Then \mathbf{Q}_{scn} interact with BEV features through self attention, further expanding the global receptive field and propagating visible features from nearby areas to distant and unseen regions beyond the viewpoint, as expressed by:

$$\mathbf{Q}_{\text{scn}} = \text{SelfAttn}(\mathbf{Q}_{\text{scn}}) \quad (7)$$

Finally, the queries are combined into BEV features based on the positions of the reference points, and C_{scn} is obtained through upsampling layers.

3.4. Feature Fusion and Losses

Feature Fusion. Prior BEV-based 3D reconstruction methods [10, 49] suffer from two limitations: (1) oversimplified height features and (2) coupled instance-scene feature interactions. When instance-class objects (e.g., pedestrians) and scene-class elements (e.g., roads) project onto the same BEV grid, their significant height variations create learning conflicts. We address this through category-decoupled height prediction:

$$V = (C_{\text{ins}} \otimes H_{\text{ins}}) + (C_{\text{scn}} \otimes H_{\text{scn}}) \quad (8)$$

where $H_{\text{ins}}, H_{\text{scn}} \in [0, 1]^{X \times Y \times Z}$ are height distributions predicted via convolutional networks, and \otimes denotes broadcasted element-wise multiplication.

We enhance V using a Local and Global Aggregator with Dynamic Fusion to produce the final 3D feature volume. Unlike CGFormer [50], only the left and front views are input to the Global Aggregator, as the BEV plane already contains sufficient detailed features for aggregation.

Training Loss. Following prior works, we use the Scene-Class Affinity Loss $\mathcal{L}_{\text{scal}}$ from Monoscene [2] for both semantic and geometric predictions, while simultaneously optimizing precision, recall, and specificity. Additionally, a class-frequency weighted cross-entropy loss \mathcal{L}_{ce} is applied, with the total loss for this part given by:

$$\mathcal{L}_{\text{ssc}} = \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{scal}}^{\text{sem}} + \mathcal{L}_{\text{ce}} \quad (9)$$

Additionally, we introduce segmentation and height prediction losses for both instance and scene categories on the BEV plane to facilitate discriminative feature learning. Segmentation loss combines cross-entropy and dice loss [34], while height prediction uses focal loss [33]:

$$\mathcal{L}_{\text{aug}} = \mathcal{L}_{\text{seg}} + \lambda_h \mathcal{L}_{\text{height}} \quad (10)$$

The \mathcal{L}_{aug} loss is applied at each decoder layer, halved for layers except the final one. We also use explicit depth loss \mathcal{L}_d [26] to supervise the depth prediction. The total loss is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{ssc}} + \lambda_2 \mathcal{L}_{\text{aug}} + \lambda_d \mathcal{L}_d \quad (11)$$

In practice, we set the weights as: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_h = 5$ and $\lambda_d = 0.01$.

4. Experiments

4.1. Dataset and Metric

DISC is evaluated on two widely used datasets: SemanticKITTI [1] and SSCBench-KITTI-360 [24]. For evaluation metrics, we employ standard measures such as Intersection over Union (IoU) and mean IoU (mIoU) for voxel-wise predictions. Additionally, we introduce instance mean

Method	IoU	InsM	ScnM	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign	
					(15.30%)	(11.15%)	(1.12%)	(0.56%)	(4.17%)	(3.92%)	(0.16%)	(0.03%)	(0.03%)	(0.20%)	(0.35%)	(0.51%)	(9.17%)	(0.27%)	(0.07%)	(0.05%)	(1.90%)	(0.20%)	(0.08%)	
<i>Temporal Inputs Methods</i>																								
VoxFormer-T [25]	43.21	4.79	22.97	13.41	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	<u>1.60</u>	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70	
HASSC-T [39]	42.87	5.27	24.51	14.38	55.30	29.60	25.90	11.30	23.10	23.00	2.90	1.90	1.50	4.90	24.80	9.80	<u>26.50</u>	<u>1.40</u>	<u>3.00</u>	0.00	14.30	7.00	7.10	
H2GFormer-T [41]	<u>43.52</u>	<u>5.37</u>	<u>25.06</u>	<u>14.60</u>	<u>57.90</u>	<u>30.40</u>	<u>30.00</u>	6.90	<u>24.00</u>	<u>23.70</u>	<u>5.20</u>	0.60	1.20	<u>5.00</u>	<u>25.20</u>	<u>10.70</u>	25.80	1.10	0.10	0.00	<u>14.60</u>	<u>7.50</u>	9.30	
HTCL [18]	44.23	6.48	28.86	17.09	64.40	34.80	33.80	12.40	25.90	27.30	5.70	<u>1.80</u>	2.20	5.40	25.30	10.80	31.20	1.10	3.10	0.90	21.10	9.00	<u>8.30</u>	
<i>Single-frame Inputs Methods</i>																								
MonoScene* [2]	34.16	3.59	19.4	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 [12]	34.25	3.49	19.88	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	
SurroundOcc [42]	34.72	3.90	20.7	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40	
OccFormer [52]	34.53	4.02	21.53	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	
IAMSSC [43]	43.74	4.50	21.12	12.37	54.00	25.50	24.70	6.90	19.20	21.30	3.80	1.10	0.60	3.90	22.70	5.80	19.40	1.50	2.90	0.50	11.90	5.30	4.10	
VoxFormer-S [25]	42.95	4.39	20.89	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	
DepthSSC [46]	<u>44.58</u>	4.87	22.26	13.11	55.64	27.25	25.72	5.78	20.46	21.94	3.74	1.35	0.98	4.17	23.37	7.64	21.56	1.34	2.79	0.28	12.94	5.87	6.23	
Symphonize [13]	42.19	6.14	24.93	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	
HASSC-S [39]	43.40	5.12	22.46	13.34	54.60	27.70	23.80	6.20	21.10	22.80	4.70	1.60	1.00	3.90	23.80	8.50	23.30	1.60	<u>4.00</u>	0.30	13.10	5.80	5.50	
StereoScene [17]	43.34	5.73	26.04	15.36	61.90	31.20	30.70	10.70	24.20	22.80	2.80	3.40	2.40	<u>6.10</u>	23.80	8.40	<u>27.00</u>	<u>2.90</u>	2.20	0.50	16.50	7.00	7.20	
H2GFormer-S [41]	44.20	5.05	23.33	13.72	56.40	28.60	26.50	4.90	22.80	23.40	4.80	0.80	0.90	4.10	24.60	9.10	23.80	1.20	2.50	0.10	13.30	6.40	6.30	
MonoOcc-S [53]	-	5.36	23.14	13.80	55.20	27.80	25.10	9.70	21.40	23.20	5.20	2.20	1.50	5.40	24.00	8.70	23.00	1.70	2.00	0.20	13.40	5.80	6.40	
MonoOcc-L [53]	-	6.95	25.29	15.63	59.10	30.90	27.10	9.80	22.90	23.90	7.20	4.50	2.40	7.70	25.00	9.80	<u>26.10</u>	2.80	4.70	0.60	16.90	7.30	8.40	
VPOcc [16]	44.09	6.03	26.31	15.65	59.10	32.30	30.90	9.70	<u>26.30</u>	24.40	5.30	3.30	<u>3.20</u>	5.60	<u>25.90</u>	9.70	25.70	2.40	2.90	0.30	17.20	6.60	6.30	
CGFormer [50]	44.41	<u>6.15</u>	<u>28.24</u>	<u>16.63</u>	64.30	<u>34.20</u>	<u>34.10</u>	<u>12.10</u>	25.80	<u>26.10</u>	4.30	3.70	1.30	2.70	24.50	11.20	29.30	1.70	3.60	0.40	<u>18.70</u>	8.70	9.30	
DISC(ours)	45.32	7.25	28.56	17.35	<u>63.10</u>	34.70	34.60	12.60	26.60	26.70	<u>5.50</u>	<u>4.00</u>	4.70	8.10	26.50	<u>10.30</u>	29.30	2.80	2.50	<u>1.10</u>	19.30	8.40	<u>8.70</u>	

Table 1. **Quantitative results on SemanticKITTI test.** * represents the reproduced results from [12, 52]. Among all methods, the top three ranked approaches are marked as **red**, **bold**, and underlined. For single-frame methods, DISC achieves SOTA performance in mIoU, IoU, InsM, and ScnM. Notably, using only single-frame input, DISC surpasses even multi-frame SOTA methods in mIoU, IoU, and InsM.

IoU (InsM) and scene mean IoU (ScnM) metrics to assess the model’s perceptual capabilities across different categories. Further details regarding the datasets and metrics are provided in the supplementary materials.

4.2. Implementation Details

Following Symphonies [13], the ResNet-50 [9] backbone and image encoder are initialized with pre-trained MaskDINO [19] weights. Generally, categories such as car, bicycle, and traffic sign are classified as instances, while road, sidewalk, and building are categorized as scene. Detailed category definitions are provided in the supplementary materials. In our experiments, we observed that our network converges faster compared to previous works, allowing us to reduce the total training epochs to 20 which is shorter than most existing approaches. We use the AdamW [27] optimizer with an initial learning rate of 2e-4 and a weight decay of 1e-4. The learning rate is reduced by a factor of 0.1 at the 12th epoch.

4.3. Comparison with state-of-the-art

Tab. 1 presents our results on the SemanticKITTI hidden test set. DISC outperforms all competing methods in IoU, mIoU, InsM, and ScnM, achieving scores of 45.30, 17.35, 7.25, and 28.56, respectively. Notably, DISC shows sig-

nificant progress in instance categories, surpassing the existing state-of-the-art (SOTA) method by 17.9%, strongly validating the effectiveness of our instance-specific design. Additionally, DISC is the first single-frame-based method to outperform SOTA multi-frame fusion methods in IoU, mIoU, and InsM, while achieving a ScnM score only 0.30 lower, demonstrating its ability to fully exploit single-frame information. DISC achieves the best performance in most classes, such as sidewalk, parking, building, motorcycle, and other-vehicle.

We also conducted experiments on SSCBench-KITTI-360. As shown in Tab. 2, DISC achieves excellent performance with an mIoU of 20.55, IoU of 47.35, InsM of 13.47, and ScnM of 28.88. Moreover, DISC outperforms all camera-based methods and LiDAR-based methods in mIoU and InsM. This analysis further confirms the effectiveness and outstanding performance of DISC.

4.4. Ablation Studies

Analysis of architecture. Tab. 3 analyzes different architectures. The baseline can be regarded as a simplified FlashOcc [49], which consists of a 2D backbone for extracting image features, a Coarse-to-fine BEV generation module for producing reliable BEV features, a BEV encoder for further processing, and a 3D global and local en-

Method					car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-obj.
	IoU	InsM	ScnM	mIoU	(2.85%)	(6.01%)	(0.01%)	(0.06%)	(5.75%)	(0.02%)	(14.98%)	(2.14%)	(6.43%)	(2.05%)	(15.67%)	(0.96%)	(41.09%)	(7.10%)	(0.22%)	(0.06%)	(4.33%)	(0.28%)
<i>LiDAR-based methods</i>																						
SSCNet [36]	53.58	4.96	28.88	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 [31]	<u>47.35</u>	2.42	24.96	13.65	20.91	0.00	0.00	0.26	0.58	0.00	<u>62.95</u>	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
<i>Camera-based methods</i>																						
MonoScene [2]	37.87	5.22	19.41	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 [12]	40.22	5.89	21.39	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 [25]	38.76	4.89	18.93	11.81	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 [52]	40.27	6.76	22.39	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
DepthSSC [46]	40.85	7.19	21.72	14.28	21.90	<u>2.36</u>	<u>4.30</u>	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.71	5.24	3.51
Symphonies [13]	<u>44.12</u>	13.11	24.05	<u>18.58</u>	30.02	1.85	5.90	25.07	12.06	8.20	54.94	13.83	32.76	6.93	35.11	8.58	38.33	11.52	<u>14.01</u>	<u>9.57</u>	14.44	11.28
CGFormer [50]	48.07	<u>12.08</u>	28.01	20.05	<u>29.85</u>	3.42	3.96	<u>17.59</u>	<u>6.79</u>	<u>6.63</u>	63.85	<u>17.15</u>	40.72	5.53	<u>42.73</u>	<u>8.22</u>	<u>38.80</u>	<u>24.94</u>	16.24	17.45	<u>10.18</u>	<u>6.77</u>
DISC(ours)	<u>47.35</u>	13.47	<u>27.63</u>	20.55	29.41	4.64	8.27	19.24	8.51	6.74	61.88	17.56	<u>40.09</u>	<u>5.27</u>	42.53	9.24	38.76	23.05	16.73	19.51	10.32	8.21

Table 2. **Quantitative results on SSCBench-KITTI360 test.** The results for most counterparts are provided in [24]. Among all methods, the top three ranked approaches are marked as **red**, **bold**, and underlined. DISC achieves SOTA results in mIoU and InsM, while surpassing LiDAR-based methods across multiple category-specific metrics.

Method	IoU \uparrow	InsM \uparrow	ScnM \uparrow	mIoU \uparrow
Baseline	20.05	6.06	19.14	12.69
+ Instance Stream	45.12	9.79	24.73	16.86
+ Scene Stream	45.85	8.55	25.80	16.73
DISC(Ours)	45.93	8.75	26.27	17.05

Table 3. **Ablation study on architecture of DISC.** DISC achieves optimal comprehensive performance.

Method	InsM \uparrow	ScnM \uparrow	mIoU \uparrow
w/o C2FBEV	8.56	25.74	16.70
w/o Image-assisted strategy	8.44	25.75	16.64
Using a fixed threshold	8.67	26.00	16.88
DISC instance query	8.75	26.27	17.05

Table 4. **Ablation study on query generator.** We evaluate diverse query generation strategies for optimal performance.

coder for handling 3D features, similar to CGFormer [50]. On top of the baseline, adding the instance stream and scene stream individually leads to significant improvements of 3.73 InsM and 6.66 ScnM for instance and scene, respectively. When integrating both the instance and scene streams, DISC achieves a balanced improvement in instance and scene metrics and delivers optimal IoU and mIoU scores, further validating the effectiveness of our methods.

Analysis of discriminative query generator. Tab. 4 provide an ablation analysis of the Discriminative Query Generator (DQG). Tab. 4 shows that removing the Coarse-to-fine BEV generation module significantly reduces both InsM and ScnM. Additionally, eliminating the front-view-assisted candidate point localization strategy and replac-

Ins. SA	Ins.-Img. CA	Ins.-Scn. CA	InsM	mIoU
×	×	×	8.55	16.73
✓	×	×	8.62	16.69
✓	✓	×	8.73	16.98
✓	×	✓	8.69	16.86
✓	✓	✓	8.75	17.05

Table 5. **Ablation study on the Adaptive Instance Layer.** Cross-attention with image plays a crucial role in instance performance.

Scn. SA	Scn.-Img. CA	ScnM	mIoU
×	×	24.73	16.86
✓	×	25.99	17.01
×	✓	25.96	16.85
✓	✓	26.27	17.05

Table 6. **Ablation study on the Global Scene Layer.** Self-attention proves critical for scene reconstruction.

ing the grid-based candidate point selection strategy with a fixed threshold has a more pronounced impact on instance than on scene. Furthermore, we show that a patch size of 4 for scene query initialization yields the best mIoU and ScnM (see the appendix for details). These results highlight the effectiveness of the DQG in providing sufficient geometric and semantic priors for query initialization, leading to optimal network performance.

Analysis of dual-attention class decoder. To analyze the flow of instance and scene queries within their respective decoder layers, we evaluate the interactions between various modules in Adaptive Instance Layer (AIL) and the Global Scene Layer (GSL). Tab. 5 shows that removing the instance-image cross-attention has the most significant im-

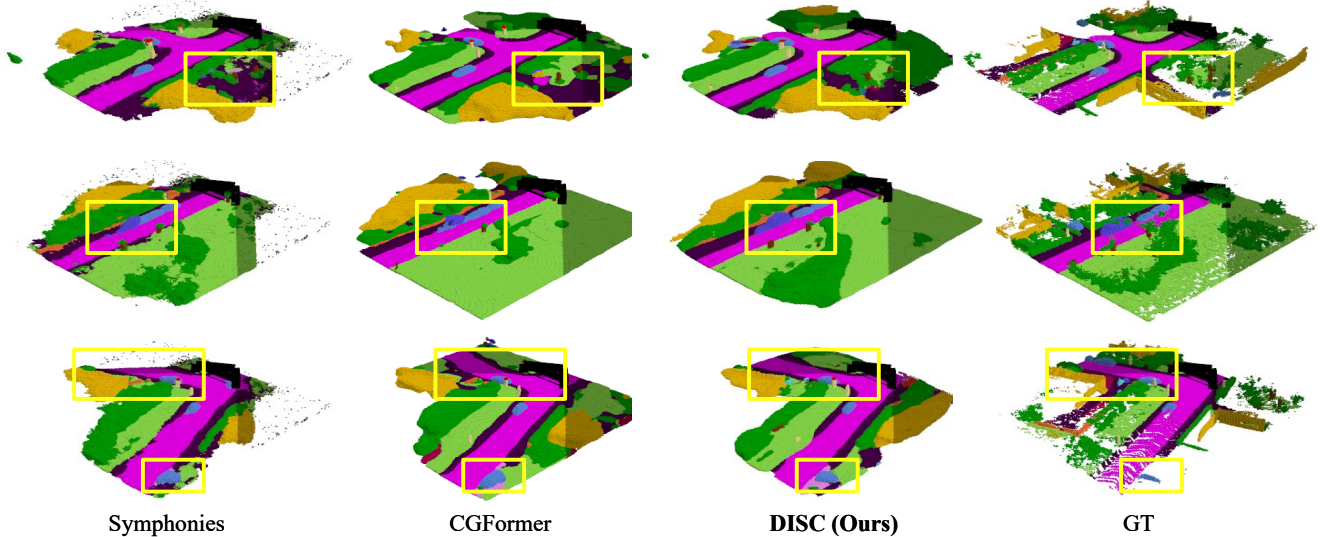


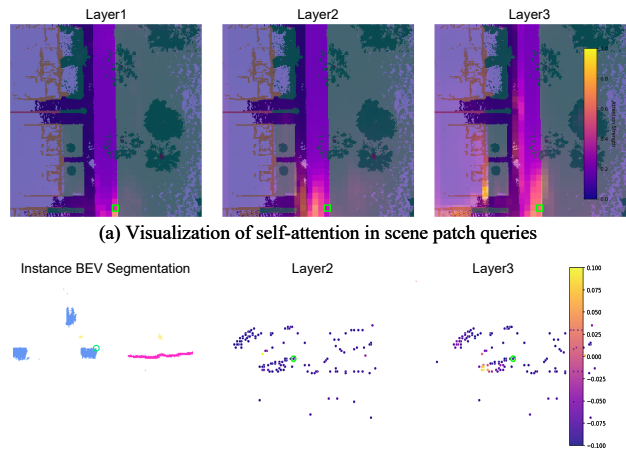
Figure 5. **Qualitative visualization results on the SemanticKITTI [1] validation set.** Compared to state-of-the-art (SOTA) methods, DISC produces more logical scene layouts and more accurate and detailed instance predictions.

impact on performance, underscoring the importance of fusing instance queries with front-view features. Tab. 6 reveals that scene categories rely more on the self attention module to refine the global scene layout compared to instance categories, highlighting the distinct reconstruction challenges faced by each category.

4.5. Visualizations

Qualitative Results. Fig. 5 visualizes predicted results on the SemanticKITTI validation set from Symphonies [13], CGFormer [50], and our proposed DISC. We compare the performance differences between DISC and other SOTA methods on scene and instance categories. As highlighted by the yellow boxes, DISC predicts more logical scene layouts (e.g., roads do not intersect with terrain) and better geometric distributions (e.g., road structures are more complete and realistic), leveraging the scene features’ global perception capabilities. For instances, DISC mitigates the impact of projection errors and occlusions, resulting in more accurate and detailed instance predictions. Additional qualitative results can be found in the Appendix.

Attention Map Analysis. In Fig. 6, we visualize the interaction mechanisms of scene and instance queries within their respective layers. The patch-based scene queries expand their receptive field through multi-layer self-attention. Fig. 6 (b) shows the ground truth distribution of instance categories and the instance query candidate points generated by the Discriminative Query Generator, which are concentrated around the ground truth positions. Additionally, in instance self-attention, queries of the same category receive higher attention weights. These results highlight the effectiveness of our discriminative processing for scene and instance categories.



(a) Visualization of self-attention in scene patch queries
 (b) Visualization of self-attention in instance point queries
 Figure 6. **Analysis of attention maps within DISC.**

5. Conclusion

In this paper, we propose **DISC**, a dual-stream neural architecture that resolves semantic completion challenges through disentangled refinement of instance and scene representations on the BEV plane. The framework incorporates an efficient query generator (**DQG**) that fuses geometric-semantic features to enhance both instance and scene queries, complemented by a dual-attention class decoder (**DACD**) comprising a global-aware scene layer for contextual ambiguity reduction and an adaptive instance layer employing dynamic feature aggregation to address occlusion and projection errors. Experimental validation demonstrates that DISC achieves **state-of-the-art** performance on SemanticKITTI and SSCBench-KITTI-360 benchmarks while maintaining computational efficiency.

References

- [1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.
- [2] 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*, pages 3991–4001, 2022.
- [3] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pages 550–567. Springer, 2022.
- [4] Sijia Chen, En Yu, Jinyang Li, and Wenbing Tao. Delving into the trajectory long-tail distribution for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19341–19351, 2024.
- [5] Sijia Chen, En Yu, and Wenbing Tao. Cross-view referring multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2204–2211, 2025.
- [6] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021.
- [7] Shi Gong, Xiaoqing Ye, Xiao Tan, Jingdong Wang, Errui Ding, Yu Zhou, and Xiang Bai. Gitnet: Geometric prior-based transformation for birds-eye-view segmentation. In *European Conference on Computer Vision*, pages 396–411. Springer, 2022.
- [8] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023.
- [9] 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*, pages 770–778, 2016.
- [10] Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird’s-eye view and perspective view. *arXiv preprint arXiv:2403.02710*, 2024.
- [11] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhao Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.
- [12] 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*, pages 9223–9232, 2023.
- [13] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024.
- [14] 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*, pages 1042–1050, 2023.
- [15] Bu Jin, Yupeng Zheng, Pengfei Li, Weize Li, Yuhang Zheng, Sujie Hu, Xinyu Liu, Jinwei Zhu, Zhijie Yan, Haiyang Sun, et al. Tod3cap: Towards 3d dense captioning in outdoor scenes. In *European Conference on Computer Vision*, pages 367–384. Springer, 2024.
- [16] Junsu Kim, Junhee Lee, Ukcheol Shin, Jean Oh, and Kyungdon Joo. Vpocc: Exploiting vanishing point for monocular 3d semantic occupancy prediction. *arXiv preprint arXiv:2408.03551*, 2024.
- [17] Bohan Li, Yasheng Sun, Xin Jin, Wenjun Zeng, Zheng Zhu, Xiaofeng Wang, Yunpeng Zhang, James Okae, Hang Xiao, and Dalong Du. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*, 1(3):6, 2023.
- [18] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pages 131–148. Springer, 2025.
- [19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.
- [20] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhao Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird’s-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [21] Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. Ovtr: End-to-end open-vocabulary multiple object tracking with transformer. In *The Thirteenth International Conference on Learning Representations*.
- [22] 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*, pages 1486–1494, 2023.
- [23] 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*, pages 1477–1485, 2023.

- [24] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Ssbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023.
- [25] 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*, pages 9087–9098, 2023.
- [26] Zhiqi Li, Wenhao 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*, pages 1–18. Springer, 2022.
- [27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [28] 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.
- [29] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.
- [30] 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*, pages 8555–8564, 2021.
- [31] 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.
- [32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.
- [33] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.
- [34] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International conference on robotics and automation (ICRA)*, pages 9200–9206. IEEE, 2022.
- [35] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2417–2426, 2022.
- [36] 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 conference on computer vision and pattern recognition*, pages 1746–1754, 2017.
- [37] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.
- [38] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024.
- [39] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024.
- [40] Wenjie Wang, Yehao Lu, Guangcong Zheng, Shuigen Zhan, Xiaoqing Ye, Zichang Tan, Jingdong Wang, Gaoang Wang, and Xi Li. Bevspread: Spread voxel pooling for bird’s-eye-view representation in vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14718–14727, 2024.
- [41] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024.
- [42] 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*, pages 21729–21740, 2023.
- [43] Haihong Xiao, Hongbin Xu, Wenxiong Kang, and Yuqiong Li. Instance-aware monocular 3d semantic scene completion. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [44] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022.
- [45] 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*, pages 3101–3109, 2021.
- [46] Jiawei Yao and Jusheng Zhang. Depthssc: Depth-spatial alignment and dynamic voxel resolution for monocular 3d semantic scene completion. *arXiv preprint arXiv:2311.17084*, 2023.
- [47] En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023.
- [48] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xi-

- angyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *European Conference on Computer Vision*, pages 425–443. Springer, 2024.
- [49] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zong-dai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023.
- [50] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, pages 1531–1555, 2024.
- [51] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird’s-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3348–3357, 2023.
- [52] 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*, pages 9433–9443, 2023.
- [53] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18398–18405. IEEE, 2024.