

Towards 3D Object-Centric Feature Learning for Semantic Scene Completion

Weihua Wang^{1,2*}, Yubo Cui^{1*}, Xiangru Lin⁴, Zhiheng Li¹, Zheng Fang^{1,2,3†}

¹Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Liaoning, China

²National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, Liaoning, China

³The Key Laboratory of Data Analytics and Optimization for Smart Industry(Northeastern University), China

⁴The University of Hong Kong

{wangweihua,ybcui21, zhli24}@stumail.neu.edu.cn, xiangru.lin@connect.hku.hk, fangzheng@mail.neu.edu.cn

Abstract

Vision-based 3D Semantic Scene Completion (SSC) has received growing attention due to its potential in autonomous driving. While most existing approaches follow an ego-centric paradigm by aggregating and diffusing features over the entire scene, they often overlook fine-grained object-level details, leading to semantic and geometric ambiguities, especially in complex environments. To address this limitation, we propose Ocean, an object-centric prediction framework that decomposes the scene into individual object instances to enable more accurate semantic occupancy prediction. Specifically, we first employ a lightweight segmentation model, MobileSAM, to extract instance masks from the input image. Then, we introduce a 3D Semantic Group Attention module that leverages linear attention to aggregate object-centric features in 3D space. To handle segmentation errors and missing instances, we further design a Global Similarity-Guided Attention module that leverages segmentation features for global interaction. Finally, we propose an Instance-aware Local Diffusion module that improves instance features through a generative process and subsequently refines the scene representation in the BEV space. Extensive experiments on the SemanticKITTI and SSCBench-KITTI360 benchmarks demonstrate that Ocean achieves state-of-the-art performance, with mIoU scores of 17.40 and 20.28, respectively.

Introduction

Semantic Scene Completion (SSC), also known as semantic occupancy prediction, has made significant progress in recent years. By partitioning the 3D space into voxels and predicting a semantic label for each voxel, SSC generates a dense and structured representation of the 3D environment. This fine-grained and enriched semantic detail can better serve downstream planning in autonomous driving. Compared to LiDAR-based approaches, vision-based methods are gaining increasing attention due to their lower cost and accessibility. By incorporating depth estimation, monocular vision-based SSC methods can infer 3D semantic information of the surrounding scene using only RGB images.

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

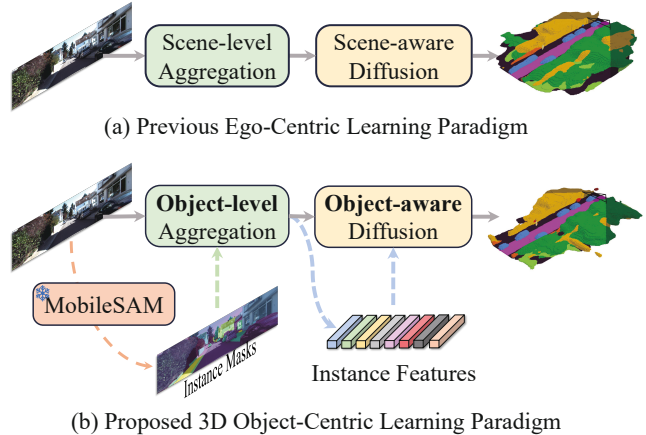


Figure 1: Comparison between our object-centric learning guided by MobileSAM and previous scene-level paradigms.

To project 2D visual features into 3D space for prediction, most previous methods follow an ego-centric paradigm. This approach uses camera-ego relations to transform 2D features into 3D features and then diffuses these features across the entire scene. For example, MonoScene (Cao and De Charette 2022) lifts the multi-scale visual feature to 3D spaces with the camera parameters and utilizes 3D convolution to complete the 3D feature to make predictions. VoxFormer (Li et al. 2023c) selects a subset of voxels as *query proposals* based on depth predictions, using cross-attention to aggregate features from the image and then employing self-attention to propagate this aggregated information to the entire 3D scene. However, despite these achievements, this global paradigm does not adequately distinguish between different voxels, resulting in unintended interactions that usually cause semantic and geometric ambiguity. Specifically, fusing features from different objects would lead to semantic confusion, while fusing features from empty and occupied areas also leads to geometric confusion. As illustrated in Figure 1(a), when multiple nearby cars are along the roadside, the empty spaces between the cars are mistakenly assigned features resembling those of the cars due to global fusion. This makes it difficult to distinguish individual cars and leads to incorrect long-trailing predictions.

In contrast, object-centric learning paradigm has been explored to emphasize more detailed feature interactions, employing instance queries (Jiang et al. 2024) or Gaussian points (Huang et al. 2024) to implicitly represent objects and aggregate features from image features. Compared to the ego-centric learning paradigm, this paradigm demonstrates greater potential in aggregating object details. However, these methods lack explicit object-level correspondences, hindering effective feature interaction at the object level and resulting in limited performance gains. Recently, large visual foundation models such as SAM (Kirillov et al. 2023) have made significant strides in visual understanding, particularly in capturing fine-grained image details. Building on this, MobileSAM (Zhang et al. 2023) adopts a lightweight design that enables the practical deployment of SAM in real-world downstream tasks. This raises the question: *Can these models be leveraged to provide finer details and enhance vision-based SSC predictions?* However, leveraging these vision foundation models poses two major challenges: 1) the mask priors are constrained to the 2D image plane, which limits their capacity for comprehensive 3D feature interaction. 2) Mistakes and omissions in the prior masks may result in noticeable performance degradation.

To address these challenges, we introduce ocean, a 3D object-centric feature learning network for semantic scene completion, which integrates a SemGroup Dual Attention (SGDA) block and an Instance-aware Local Diffusion (ILD) module. The SGDA block incorporates two key components, 3D Semantic Group Attention (SGA3D) and Global Similarity-Guided Attention (GSGA), to enhance 3D semantic understanding. Specifically, the SGA3D module facilitates object-level semantic aggregation by performing feature interactions within individual instances in 3D space, while GSGA leverages prior features to correct mask errors and reduce foreground omissions caused by imperfect masks. Furthermore, the ILD module leverages instance-level features to reconstruct bird’s-eye view representations, which are then used to refine the entire scene feature for finer-grained semantic representation.

In summary, our contributions are as follows:

- To address object-level semantic and geometric ambiguities in scenes, we propose Ocean, a novel approach that explicitly aggregating 3D object-level features and effectively enhancing scene feature representations.
- To comprehensively utilize MobileSAM priors, we propose SGDA and ILD module, which are designed to aggregate object-centric features and diffuse instance-aware information throughout the scene, respectively.
- Experiments on the SemanticKITTI and SSC-BenchmarkKITTI360 show that Ocean achieves state-of-the-art performance with mIoU of 17.40 and 20.28, respectively.

Related Works

Vision-based 3D Perception

The perception system plays a crucial role in autonomous driving. Compared to LiDAR-based methods, vision-based methods (Phillion and Fidler 2020; Li et al. 2022) offer advantages in terms of lower cost and easier deployment, thus

have garnered increasing attention. LSS (Phillion and Fidler 2020) lifts 2D image features into 3D by predicting a per-pixel depth distribution and computing its outer product with the features. However, LSS is sensitive to depth. In contrast, BEVformer (Li et al. 2022) directly defines grid-shaped BEV queries to aggregate the BEV features. FB-BEV (Li et al. 2023d) analyzes the limitations of forward and backward projection methods and proposes a unified forward-backward paradigm to address them. Nowadays, with the development of visual research, visual perception has made tremendous progress in detection (Qi et al. 2024; Zhang et al. 2025), segmentation (Lu, Tsai, and Chen 2025).

Semantic Scene Completion

MonoScene (Cao and De Charette 2022) is the first to leverage monocular images for the 3D SSC task. It lifts 2D visual features to 3D space and employs a 3D U-Net to extract voxel features for the final prediction. Moreover, TPVFormer (Huang et al. 2023) introduces a tri-perspective view (TPV) and utilizes cross-view hybrid attention to enable interaction of features across the TPV planes. OccFormer (Zhang, Zhu, and Du 2023) decomposes 3D feature processing into local and global scales to enable long-range and dynamic interactions. Voxformer (Li et al. 2023c) incorporates depth priors and adopts a sparse-to-dense approach for 3D feature interactions. Building upon these works, MonoOcc (Zheng et al. 2024) improves 3D occupancy via cross-attention and temporal distillation. CGFormer (Yu et al. 2024) addresses query ambiguity with a context and geometry aware design. LOMA (Cui et al. 2025) uses a VL-aware generator and Tri-plane Fusion to boost language integration and global 3D modeling. To reduce the heavy computation burden, SparseOcc (Liu et al. 2024) decouples semantics and geometry, introducing a sparse and efficient framework. OctreeOcc (Lu et al. 2024) addresses dense-grid limitations by adaptively utilizing octree representations. GaussianFormer (Huang et al. 2024) first employs sparse gaussian representations for feature aggregation. Different from the above methods, Symphonies (Jiang et al. 2024) uses learnable instance queries to iteratively aggregate 3D features in an object-centric pipeline. This paper adopts object-centric feature learning but uniquely leverages object priors for explicit object-level interactions.

Object-Centric Learning

Object-centric learning aims to represent the entire scene by utilizing multiple distinct objects within it. AIR (Eslami et al. 2016) models each object using appearance, position, and presence, and these components are combined to form the final scene. In contrast, MONET (Burgess et al. 2019) use spatial mixture models for scene representation, but are computationally inefficient. To solve this, Slot Attention (Locatello et al. 2020) develops a grouping strategy to create distinct slot representations. However, slot counts can’t handle varying object numbers. AdaSlot (Fan et al. 2024) dynamically adjusts slots based on instance complexity. Inspired by slot-based methods, we propose an instance-aware feature diffusion for 3D object-centric learning.

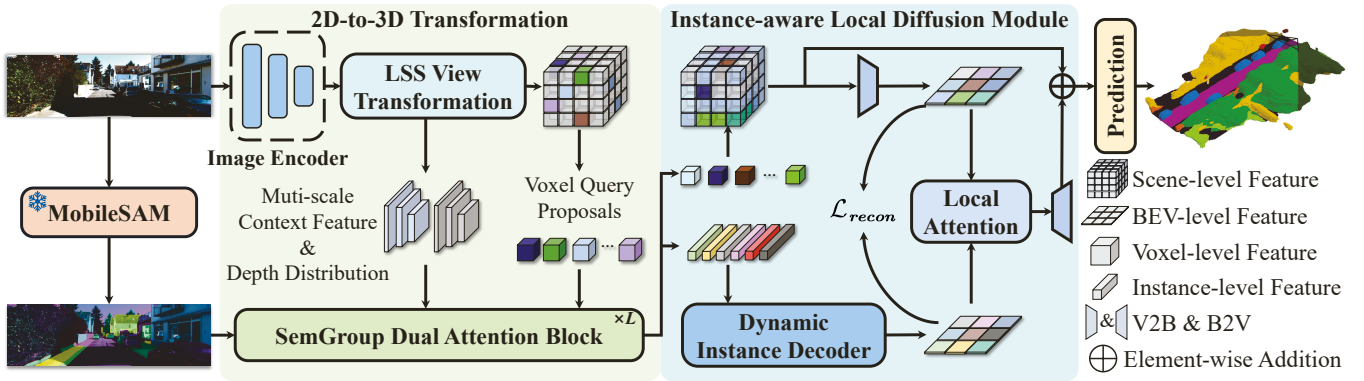


Figure 2: Overview of the proposed Ocean architecture. Given the monocular image as input, we first extract visual features using an image encoder and lift them into 3D space following LSS. To enable object-centric feature learning, we segment the scene using MobileSAM and design the SGDA block to aggregate features through both local and global attention. Furthermore, we propose the ILD module to refine the overall scene representation by incorporating instance-level features.

Methods

Overview

The overall architecture of the proposed method is shown in Figure 2. Given a monocular image $\mathcal{I} \in \mathbb{R}^{3 \times H \times W}$, the goal of camera-based 3D occupancy prediction is to reconstruct a 3D occupancy scene $\mathcal{O} \in \mathbb{R}^{X \times Y \times Z \times (M+1)}$, where M denotes the number of semantic classes. First of all, we employ an image encoder to extract multi-scale visual features $\{\mathcal{F}^{(s)} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_1^{(s)}}\}$ at three different downsampling ratios $s \in \{4, 8, 16\}$. Then, each scale-specific feature $\mathcal{F}^{(s)}$ is further processed by two separate modules to predict the corresponding depth distribution $\mathcal{D}^{(s)} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times D}$ and context feature $\mathcal{X}^{(s)} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_2}$, where D is the number of discretized depth bins. We compute the outer product between the predicted depth distribution \mathcal{D} and the context feature \mathcal{X} , and then apply voxel pooling (Pillion and Fidler 2020) to lift the 2D image features into a 3D voxel representation $\mathcal{V} \in \mathbb{R}^{x \times y \times z \times C_2}$, where x, y, z denote the spatial dimensions of the 3D scene. This lifting operation is performed using the features at scale $s = 8$, which provides a good trade-off between spatial resolution and computational efficiency. Following previous works (Li et al. 2023c; Yu et al. 2024), we also use a pre-trained network (Shamsafar et al. 2022) to predict the depth map, and then select the 3D voxel queries $\mathcal{Q} \in \mathbb{R}^{N \times C_2}$ ($0 \leq N \leq h \times w \times z$), called query proposals, based on the depth prediction as follows:

$$\mathcal{Q} = \mathcal{V}[M] \quad (1)$$

where M is a binary mask indicating whether the voxel is occupied based on the depth prediction. Subsequently, to incorporate the object-level mask guidance from MobileSAM into the SSC task, we propose the SemGroup Dual Attention (SGDA) block, which aggregates query proposals with visual features at the object level. Furthermore, we introduce the Instance-aware Local Diffusion (ILD) module, which leverages instance priors to refine scene features through localized feature fusion. Finally, the Semantic Scene Completion prediction is generated through a 3D prediction head.

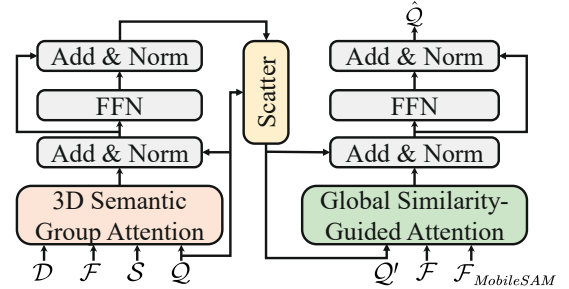


Figure 3: The details of SemGroup Dual Attention Block.

SemGroup Dual Attention Block

To mitigate the ambiguity of object-level representations introduced by the ego-centric learning paradigm, we propose a novel object-centric framework. As shown in Figure 3, our approach introduces the SGDA block, which includes a 3D Semantic Group Attention module for object-centric local aggregation and a Global Similarity-Guided Attention module for context global interaction. This dual attention mechanism effectively captures both fine-grained local details and global contextual cues, significantly enhancing the quality and completeness of scene representations.

Semantic Group Attention. Given a set of query proposals $\mathcal{Q} \in \mathbb{R}^{N \times C_2}$, our goal is to leverage visual features to enhance their representation for perception improvement. To this end, we first project the 3D query proposals onto the image plane using the camera’s intrinsic and extrinsic matrices. For a query located at (x, y, z) in the ego-centric coordinate system, its corresponding 3D pixel coordinate (u, v, d) in the image space is obtained as follows:

$$d \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot [R|t] \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (2)$$

where K is the intrinsic matrix and $[R|t]$ are the camera extrinsic matrices, d is the depth of the projected pixels.

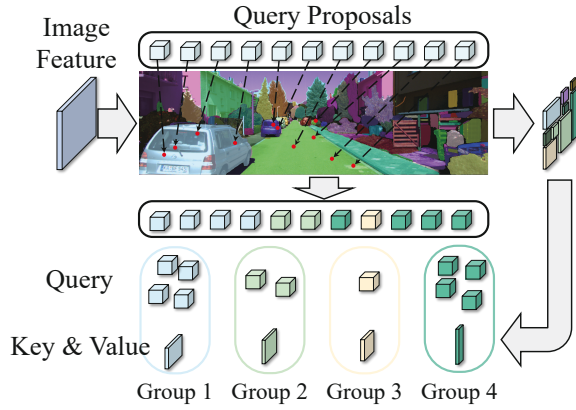


Figure 4: The Semantic Grouping. 3D query proposals are projected onto the image plane, assigned instance IDs via nearest-neighbor sampling, and clustered with image pixels of the same instance for aggregation using linear attention.

After projection, we obtain 3D proposals that are aligned with the 2D image plane. These proposals are then used to aggregate corresponding instance-level image features on the image plane, facilitating an object-centric learning paradigm. However, acquiring accurate instance information from the image is inherently challenging, and learning features between image features of the same instance and their corresponding proposals remains a non-trivial task.

To this end, we adopt MobileSAM to extract per-pixel instance masks with high efficiency. As illustrated in Figure 4, MobileSAM labels each pixel with an instance ID, generating a mask $\mathcal{S} \in \mathbb{R}^{H \times W \times 1}$ where pixels labeled as 0 denote background. Each 3D projection point is mapped to the image plane and assigned the instance ID of its nearest pixel. We use the nearest-neighbor approach to assign an instance ID to each projection point based on its projected coordinates (u, v) to group the projection points with pixels. Pixels and projection points sharing the same instance are grouped into the same clusters. Projection points associated with background pixels are excluded from this process.

Meanwhile, higher-level features contain rich semantic information, while lower-level features preserve finer-grained texture details. For this purpose, we first downsample the instance masks to match the resolutions of the multi-scale features. We then group features across different scales based on the downsampled masks $\{\mathcal{S}_i\}_{i=0}^s$, and apply an attention mechanism within each group, where the query proposals serve as the query Q and the grouped pixel-level features serve as the key K and value V , thereby aggregating multi-scale image features corresponding to the same instance ID. To further improve speed and memory efficiency, we employ scattered linear attention (He et al. 2024) for feature aggregation, formulated as follows:

$$\tilde{Q} = \text{Concat} \left[\frac{\varphi(Q^j) \sum_{i=1}^{m^j} \varphi(K_i^j)^T V_i^j}{\varphi(Q^j) \sum_{i=1}^{m^j} \varphi(K_i^j)^T} \right]_{j=1}^M \quad (3)$$

where M denotes the number of instances, m^j is the number of pixels in the cluster, and $\varphi(\cdot)$ denotes the kernel function.

3D Extension with Depth Similarity. With the proposed SGA, we integrate the MobileSAM into SSC and perform object-centric learning. However, this integration can only fuse features in 2D image space, limiting the model’s ability to perceive features in 3D space, especially geometric perception. Meanwhile, due to the 2D limitation, the prior could not guide the 3D feature diffusion to capture detailed object information. Thus, lifting the 2D prior to 3D space is essential for enhancing both semantic and geometric perception.

Inspired by DFA3D (Li et al. 2023a), we extend SGA to 3D space by integrating depth information to enhance feature aggregation, particularly to capture richer geometric details. Specifically, within each cluster, the depth of grouped pixels is obtained from the predicted depth map $\mathcal{D}^{(s)}$, while the depth of grouped proposals is computed according to Equation 2. However, pixel depth is represented as an uncertain probability distribution, whereas proposal depth is a deterministic computed value. This discrepancy hinders direct similarity computation between them.

Luckily, we observe that for each pixel, its predicted depth is represented by a softmax probability over multiple depth bins within a given depth range, which can also be interpreted as the pixel’s similarity to each depth bin. Meanwhile, we can also determine the corresponding depth bin for each projected depth of the proposal. Therefore, we can use the depth bin as a bridge to generate the depth similarity between the proposals and the pixels. Specifically, for a cluster of n pixels and m proposals, we first gather the depth distribution of the pixels from the depth prediction. Next, for each proposal, we compute its projected depth value and gather the corresponding depth bin. We then extract the corresponding probability values from the pixel depth distribution, resulting in an $m \times n$ depth similarity matrix. This approach not only effectively represents the depth similarity between proposals and pixels but also fully leverages the predicted \mathcal{D} . Finally, we achieve the SGA3D as follows:

$$\tilde{Q} = \text{Concat} \left[\frac{\varphi(Q^j) A^j \sum_{i=1}^{m^j} \varphi(K_i^j)^T V_i^j}{\varphi(Q^j) \sum_{i=1}^{m^j} \varphi(K_i^j)^T} \right]_{j=1}^M \quad (4)$$

where A^j represents the depth similarity in the j -th cluster.

Global Similarity-Guided Attention. Benefiting from the SGA3D module, our method enables instance-level feature aggregation in 3D space. However, SGA3D focuses on local feature interactions within individual instances, and the masks generated by MobileSAM may suffer from inaccuracies and missing instances. These limitations can significantly impact the overall prediction performance.

To address these limitations while maintaining the object-centric paradigm, we introduce the Global Similarity-Guided Attention (GSGA) module. It dynamically aggregates global image features to query proposals via a deformable attention mechanism, guided by global features from MobileSAM to reduce the impact of inaccurate masks. Specifically, we compute the similarity between each query proposal and the MobileSAM intermediate features sampled at deformable offsets, and use these scores to filter and emphasize features belonging to the same object instance. This

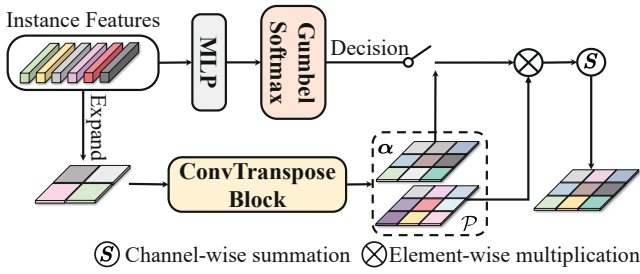


Figure 5: The details of the Dynamic Instance Decoder. Given the instance features, we reconstruct them into the scene-level BEV representation using a transposed convolutional decoder. Furthermore, we employ Gumbel Softmax to enable dynamic instance selection.

instance-aware guidance improves the relevance of feature aggregation, while the deformable mechanism enables the model to flexibly attend to informative regions beyond rigid segmentation boundaries. The final aggregated representation for each proposal is computed as:

$$\hat{Q} = \sum_k (G_k W \mathcal{F}(p_q + \Delta p_k)) \odot A_k \quad (5)$$

Here, \odot denotes element-wise multiplication, k is the number of sampling points, G_k denotes the similarity matrix between the query and the MobileSAM feature, A_k represents the learnable attention weight at the k -th sampling location for a given query, and Δp_k is the offset applied to the query position p_q . W denotes the projection weight matrix.

Instance-aware Local Diffusion Module

Although the proposed SGDA block allows 3D query proposals to capture rich image features, many voxels still lack sufficient semantic information due to projection limitations. To address this, we introduce an Instance-aware Local Diffusion module that enhances spatial consistency by propagating features across the scene.

Dynamic Instance Decoder. To effectively leverage prior masks while preserving the instance-centric design, we sum grouped features across multiple scales based on instance masks to obtain aggregated instance-level representations. This simple summation incurs minimal computational cost and, due to its sensitivity to instance size, effectively preserves spatial characteristics at the instance level.

Given the extracted instance features, we further reconstruct corresponding BEV representations to enhance spatial and semantic understanding. However, deriving fine-grained BEV features from independent instance representations is still a challenging task in complex and dynamic scenarios. To address this, we employ a generative strategy, using the voxel features from SGDA as guidance to dynamically generate BEV features enriched with instance information, as illustrated in Figure 2. Specifically, we employ a lightweight deconvolution block to generate features with a dimension of $x \times y \times (C_2 + 1)$ from the instance features. As shown in Figure 5, we also use a two-layer MLP along with Gumbel-Softmax to predict the one-hot decision \mathcal{Z} for each instance

feature. Finally, we apply Softmax to normalize α based on \mathcal{Z} , then combine the weighted BEV features to reconstruct the scene BEV representation. The process could be represented as follows:

$$w_l = \frac{\exp(\alpha_l)}{\sum_{l=1}^L \exp(\alpha_l)}, \hat{w}_l = \frac{\mathcal{Z}_l w_l}{\sum_{l=1}^L \mathcal{Z}_l w_l + \epsilon} \quad (6)$$

$$\hat{\mathcal{P}} = \sum_{l=1}^L \mathcal{P}_l \odot \hat{w}_l \quad (7)$$

where L is the number of instances, ϵ is a small positive value for stable computation, \mathcal{P}_i is i -th BEV feature. Furthermore, inspired by the Slot-based method (Fan et al. 2024), we introduce a reconstruction loss to constrain the generated BEV features, enhancing their quality and accelerating convergence. The loss function is defined as follows:

$$\mathcal{L}_{recon} = \sum_{i=1}^N \left\| \hat{\mathcal{P}}^{(i)} - \mathcal{F}_{bev}^{(i)} \right\|_2^2 \quad (8)$$

where N represents the total number of elements, and $\mathcal{F}_{bev}^{(i)}$ denotes the feature corresponding to the Bird's Eye View (BEV) representation of the entire scene.

Local Attention Refinement. Given the instance-aware BEV features, we use them to refine the aggregated BEV features. Typically, neighboring voxels contain similar features. Therefore, we use window attention (Liu et al. 2021) to refine the BEV features in a local manner. Specifically, we take the aggregated features as the query Q and the instance-aware features as the key K and value V . Finally, we use a 2D convolution layer to convert the refined BEV feature back to 3D shape and make predictions.

Experiments

Implementation Details

Following previous works (Cao and De Charette 2022), we use EfficientNetB7 (Tan and Le 2019) as our 2D backbone. The SemGroup Dual Attention block consists of 3 layers. The query proposals from SGDA are scattered back to a 3D feature of size $128 \times 128 \times 16$ with 128 channels, which is then processed by the ILD. We train Ocean on 4 GeForce 3090 GPUs, using the AdamW optimizer (Loshchilov 2017) with an initial learning rate of $3e-4$ and weight decay of 0.01.

Comparisons with the State-of-the-Art Methods

As shown in Table 1 and Table 2, our method achieves the best performance on both the SemanticKITTI (Behley et al. 2019) and the SSCBench-KITTI360 (Li et al. 2023b) datasets. On the SemanticKITTI dataset, Ocean reaches an IoU of 46.40 and a mIoU of 17.39, outperforming CGFormer by 1.21 and 0.77, respectively. It also surpasses HTCL (Li et al. 2024), which utilizes temporal information, demonstrating the effectiveness of our instance-level feature aggregation. Additionally, the SemanticKITTI validation visualizations in Figure 6 clearly demonstrate that our method achieves superior capability in capturing fine-grained scene structure, especially for objects with strong edge or contour

Method																					
	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
MonoScene	34.16	11.08	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1
TPVFormer	34.25	11.26	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5
VoxFormer	42.95	12.20	53.9	25.3	21.1	5.6	19.8	20.8	3.5	1.0	0.7	3.7	22.4	7.5	21.3	1.4	2.6	0.2	11.1	5.1	4.9
OccFormer	34.53	12.32	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7
MonoOcc	-	13.80	55.2	27.8	25.1	9.7	21.4	23.2	<u>5.2</u>	2.2	1.5	5.4	24.0	8.7	23.0	1.7	2.0	0.2	13.4	5.8	6.4
Symphonies	42.19	15.04	58.4	29.3	26.9	11.7	24.7	23.6	3.2	3.6	2.6	<u>5.6</u>	24.2	10.0	23.1	3.2	1.9	2.0	16.1	7.7	8.0
LOMA	43.01	15.10	58.0	31.8	32.2	9.5	25.3	24.9	4.1	1.7	1.7	6.4	25.6	8.7	24.7	1.4	1.7	0.6	16.8	6.5	6.1
CGFormer	<u>44.41</u>	16.63	64.3	34.2	34.1	12.1	25.8	26.1	4.3	<u>3.7</u>	1.3	2.7	24.5	11.2	29.3	1.7	3.6	0.4	18.7	<u>8.7</u>	9.3
HTCL	44.23	<u>17.09</u>	<u>64.4</u>	<u>34.8</u>	<u>33.8</u>	<u>12.4</u>	25.9	27.3	5.7	<u>1.8</u>	<u>2.2</u>	5.4	25.3	10.8	31.2	1.1	3.1	0.9	21.1	9.0	8.3
Ocean (Ours)	45.62	17.40	65.1	34.9	33.7	12.8	25.4	<u>26.8</u>	<u>5.2</u>	4.9	1.9	<u>5.6</u>	26.9	11.6	<u>30.7</u>	<u>2.3</u>	2.2	<u>1.7</u>	<u>20.8</u>	<u>8.7</u>	9.5

Table 1: Quantitative results on SemanticKITTI test set. The best and second results are in **bold** and underlined, respectively.

Method																				
	IoU	mIoU	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traffic-sign	other-struct.	other-obj.
MonoScene	37.87	12.31	19.3	0.4	0.6	8.0	2.0	0.9	48.4	11.4	28.1	3.3	32.9	3.5	26.2	16.8	6.9	5.7	4.2	3.1
TPVFormer	40.22	13.64	21.6	1.1	1.4	8.2	2.6	2.4	53.0	12.0	31.1	3.8	34.8	4.8	30.1	17.5	7.5	5.9	5.5	2.7
VoxFormer	38.76	11.91	17.8	1.2	0.9	4.6	2.1	1.6	47.0	9.7	27.2	2.9	31.2	5.0	29.0	14.7	6.5	6.9	3.8	2.4
OccFormer	40.27	13.81	22.6	0.7	0.3	9.9	3.8	2.8	54.3	13.4	31.5	3.6	36.4	4.8	31.0	19.5	7.8	8.5	7.0	4.6
Symphonies	44.12	18.58	30.0	1.9	5.9	25.1	12.1	8.2	54.9	13.8	32.8	6.9	35.1	<u>8.6</u>	38.3	11.5	14.0	9.6	14.4	11.3
CGFormer	<u>48.07</u>	<u>20.05</u>	<u>29.9</u>	<u>3.4</u>	4.0	<u>17.6</u>	6.8	6.7	63.9	17.2	<u>40.7</u>	<u>5.5</u>	42.7	8.2	38.8	24.9	<u>16.2</u>	<u>17.5</u>	10.2	6.8
SGFormer	46.35	18.30	27.8	0.9	2.6	10.7	5.7	4.3	61.0	13.2	37.0	5.1	<u>43.1</u>	7.5	<u>39.0</u>	24.9	15.8	16.9	8.9	5.3
Ocean (Ours)	48.19	20.28	29.3	3.7	<u>4.6</u>	15.1	<u>7.7</u>	<u>6.8</u>	<u>63.7</u>	<u>17.0</u>	40.9	5.0	43.7	8.9	39.2	<u>24.7</u>	16.7	19.2	<u>10.8</u>	<u>8.3</u>

Table 2: Quantitative results on SSCBench-KITTI360 test set. The best and second results are in **bold** and underlined.

cues, such as cars and roads. On the SSCBench-KITTI360 dataset, Ocean exceeds CGFormer by 0.12 in IoU and 0.23 in mIoU, and further outperforms SGFormer (Guo et al. 2025), which leverages additional remote sensing modalities. These results confirm the advantages of our object-centric 3D scene understanding approach. On both datasets, Ocean achieves outstanding performance in recognizing objects with consistent surface information, such as the road, sidewalks, and even smaller items such as poles and traffic signals, thanks to the guidance of prior knowledge and the design of the SGA3D algorithm. For objects with more complex surface information, like bicycles and vegetation, which are difficult for segmentation algorithms to distinguish, Ocean also delivers superior results, as the GSGA algorithm reduces reliance on prior information, mitigating information loss and segmentation errors.

Ablation Studies

To analyze the effectiveness of each module, we conduct ablation studies on the SemanticKITTI validation set.

Main Modules. Table 3 presents an ablation study comparing the core modules of our method. Compared to our

	SGDA		ILD		Performance	
	SGA3D	GSGA	LA	DID	IoU \uparrow	mIoU \uparrow
M0					44.62	15.80
M1	✓				45.77	16.49
M2	✓	✓			46.01	16.80
M3	✓	✓	✓		45.39	17.10
M4	✓	✓	✓	✓	46.40	17.39

Table 3: Ablation study on each module in Ocean.

baseline, the introduction of prior knowledge and the application of SGA3D result in an improvement of 1.15/0.69 in IoU/mIoU for M1, demonstrating the effectiveness of our design. However, due to the mistakes and omissions in the prior masks, the model has not fully learned the scene information. Therefore, M2 further incorporates GSGA and achieves 46.01/16.80 in IoU/mIoU. This result demonstrates that aggregating features at the instance level feature in 3D space can effectively enhance both semantic and geometric perception. Furthermore, after incorporating local fea-

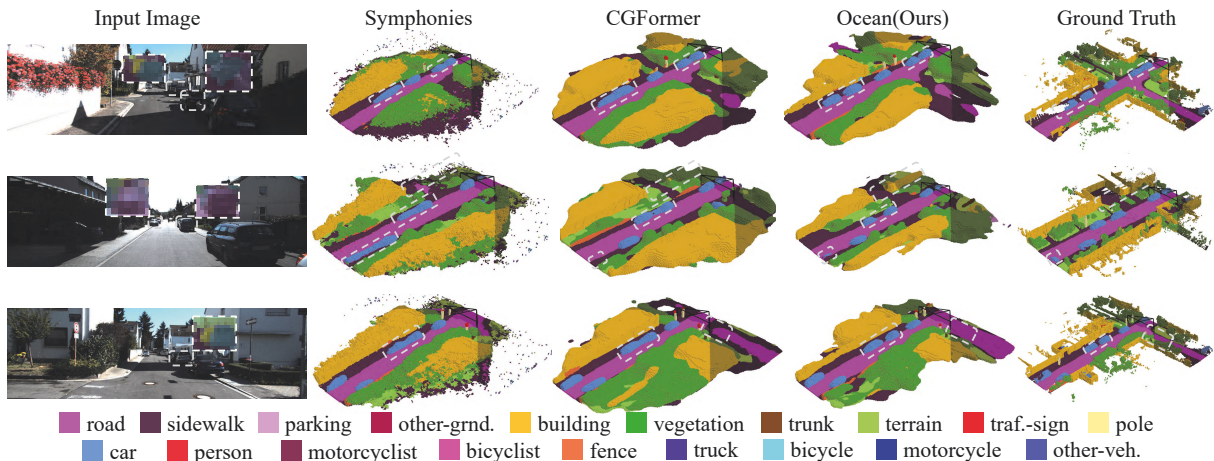


Figure 6: Qualitative visualizations on SemanticKITTI validation set.

Method	IoU \uparrow	mIoU \uparrow	Param	FLOPs
SGA3D(Ours)	46.40	17.39	0.280M	1.333G
SGA3D w/o Ms.	46.44	16.64	0.280M	1.298G
SGA3D w/o Ext.	45.84	17.30	0.280M	1.333G
DFA2D	46.14	16.39	0.202M	2.231G
DFA3D	46.15	16.97	0.330M	2.807G

Table 4: Ablation study on 3D Semantic Group Attention.

Method	IoU \uparrow	mIoU \uparrow
Full Model(Ours)	46.40	17.39
w/o GSGA	46.13	16.98
GSGA w/o Sim.	46.07	16.36

Table 5: Ablation study in Global Similarity-Guided Attention.

ture diffusion, mIoU improved by 0.3. However, due to the voxel2bev operation, which enhances the instance’s semantic information, some geometric information is inevitably lost. To address this issue, we further incorporate the Dynamic Instance Decoder (DID), which adopts a generative paradigm to enhance the overall scene semantic information and partially mitigate the loss of geometric details.

3D Semantic Group Attention. As shown in Table 4, we first evaluate the two core components of SGA3D. Without(w/o) the multi-scale image features, the model experiences a significant drop in semantic understanding, while removing the 3D extension design leads to a noticeable decline in geometric reasoning, demonstrating the distinct and indispensable contributions of both components to the effectiveness of our method. Furthermore, substituting SGA3D with DFA3D or DFA2D demonstrates the superior performance of SGA3D, which achieves mIoU gains of 1.00 and 0.42 over DFA2D and DFA3D, respectively, while maintaining a lower parameter count and computational overhead.

Method	IoU \uparrow	mIoU \uparrow
Direct Sum	45.79	16.45
Softmax-Weighted	45.89	16.88
Sigmoid-Weighted	46.30	16.92
Dynamic Selection (Ours)	46.40	17.39

Table 6: Ablation study in Dynamic Instance Decoder.

Global Similarity-Guided Attention. We conducted an ablation study on the GSGA module in Table 5. Removing GSGA leads to a decline in both IoU and mIoU metrics, indicating that it effectively mitigates the information loss caused by the foreground dependency in SGA3D and alleviates the spatial constraints imposed by the prior mask. In a separate experiment, removing similarity-guided global feature processing and replacing it with a naive global operation introduces ambiguity into both geometry and semantic features causing a degradation in overall performance. These results collectively validate the effectiveness of our design.

Dynamic Instance Decoder. In Table 6, we compare different fusion strategies in the Dynamic Instance Decoder. Compared to directly summing, the weighted approaches including sigmoid and softmax show better performance, verifying that different instances have different effects in fusion. Finally, by dynamically predicting the decision of each instance, our approach achieves the best performance.

Conclusion

We propose Ocean, an object-centric framework for vision-based 3D Semantic Scene Completion. Unlike ego-centric approaches, Ocean models scenes from an instance-level perspective to enhance semantic and geometric understanding. It leverages instance priors to guide object-centric feature learning through attention-based aggregation and spatial diffusion. Ocean achieves state-of-the-art performance on the SemanticKITTI and SSCBench-KITTI-360 datasets, showing the effectiveness of the object-centric paradigm.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62073066 and in part by 111 Project under Grant B16009.

References

- Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.
- Burgess, C. P.; Matthey, L.; Watters, N.; Kabra, R.; Higgins, I.; Botvinick, M.; and Lerchner, A. 2019. Monet: Unsupervised scene decomposition and representation. *arXiv:1901.11390*.
- Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.
- Cui, Y.; Li, Z.; Wang, J.; and Fang, Z. 2025. LOMA: Language-assisted semantic occupancy network via triplane mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2609–2617.
- Eslami, S.; Heess, N.; Weber, T.; Tassa, Y.; Szepesvari, D.; Hinton, G. E.; et al. 2016. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in neural information processing systems*, 29.
- Fan, K.; Bai, Z.; Xiao, T.; He, T.; Horn, M.; Fu, Y.; Locatello, F.; and Zhang, Z. 2024. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23062–23071.
- Guo, X.; Hu, J.; Hu, J.; Bao, H.; and Zhang, G. 2025. SGFormer: Satellite-Ground Fusion for 3D Semantic Scene Completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11929–11938.
- He, C.; Li, R.; Zhang, G.; and Zhang, L. 2024. ScatterFormer: Efficient Voxel Transformer with Scattered Linear Attention. *arXiv preprint arXiv:2401.00912*.
- Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9223–9232.
- Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2024. GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction. *arXiv:2405.17429*.
- Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Lin, T.; Liu, W.; and Wang, X. 2024. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20258–20267.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.
- Li, B.; Deng, J.; Zhang, W.; Liang, Z.; Du, D.; Jin, X.; and Zeng, W. 2024. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, 131–148. Springer.
- Li, H.; Zhang, H.; Zeng, Z.; Liu, S.; Li, F.; Ren, T.; and Zhang, L. 2023a. DFA3D: 3D Deformable Attention For 2D-to-3D Feature Lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6684–6693.
- Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; Wang, Y.; Zhao, H.; Yu, Z.; and Feng, C. 2023b. SSCBench: Monocular 3D Semantic Scene Completion Benchmark in Street Views. *arXiv:2306.09001*.
- Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023c. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9087–9098.
- Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.
- Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Alvarez, J. M. 2023d. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6919–6928.
- Liu, H.; Chen, Y.; Wang, H.; Yang, Z.; Li, T.; Zeng, J.; Chen, L.; Li, H.; and Wang, L. 2024. Fully sparse 3d occupancy prediction. In *European Conference on Computer Vision*, 54–71. Springer.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33: 11525–11538.
- Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Lu, S.-W.; Tsai, Y.-H.; and Chen, Y.-T. 2025. Toward Real-world BEV Perception: Depth Uncertainty Estimation via Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17124–17133.
- Lu, Y.; Zhu, X.; Wang, T.; and Ma, Y. 2024. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. *Advances in Neural Information Processing Systems*, 37: 79618–79641.
- Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

- Qi, Z.; Wang, J.; Wu, X.; and Zhao, H. 2024. Ocbev: Object-centric bev transformer for multi-view 3d object detection. In *2024 International Conference on 3D Vision (3DV)*, 1188–1197. IEEE.
- Shamsafar, F.; Woerz, S.; Rahim, R.; and Zell, A. 2022. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2417–2426.
- Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.
- Yu, Z.; Zhang, R.; Ying, J.; Yu, J.; Hu, X.; Luo, L.; Cao, S.-Y.; and Shen, H.-L. 2024. Context and geometry aware voxel transformer for semantic scene completion. *arXiv preprint arXiv:2405.13675*.
- Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.
- Zhang, J.; Zhang, Y.; Qi, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2025. Geobev: Learning geometric bev representation for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9960–9968.
- Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9433–9443.
- Zheng, Y.; Li, X.; Li, P.; Zheng, Y.; Jin, B.; Zhong, C.; Long, X.; Zhao, H.; and Zhang, Q. 2024. Monoocc: Digging into monocular semantic occupancy prediction. *arXiv preprint arXiv:2403.08766*.