IEEE International Conference on Systems, Man, and Cybernetics (SMC), Vienna, Austria, 2025.

OC-SOP: Enhancing Vision-Based 3D Semantic Occupancy Prediction by Object-Centric Awareness

Helin Cao and Sven Behnke

***Abstract*— Autonomous driving perception faces significant challenges due to occlusions and incomplete scene data in the environment. To overcome these issues, the task of semantic occupancy prediction (SOP) is proposed, which aims to jointly infer both the geometry and semantic labels of a scene from images. However, conventional camera-based methods typically treat all categories equally and primarily rely on local features, leading to suboptimal predictions, especially for dynamic foreground objects. To address this, we propose Object-Centric SOP (OC-SOP), a framework that integrates high-level object-centric cues extracted via a detection branch into the semantic occupancy prediction pipeline. This object-centric integration significantly enhances the prediction accuracy for foreground objects and achieves state-of-the-art performance among all categories on SemanticKITTI.**

I. Introduction

In current commercial autonomous driving frameworks, vision-only solutions primarily rely on multiple cameras mounted around the vehicle to perceive the surrounding environment. This approach not only mimics the way human drivers depend on vision but also significantly reduces system costs by eliminating expensive sensors like LiDAR. Such cost efficiency, combined with the rich contextual and high-resolution data captured by cameras, offers a promising pathway toward human-level driving capabilities and smart systems in a dynamic world. However, vision-only approaches face significant challenges in 3D scene understanding. Although state-of-the-art 2D semantic segmentation models can extract detailed pixel-wise semantic information from images, the absence of explicit depth information makes it difficult to accurately reconstruct 3D geometry and effectively map semantics into 3D space. Moreover, occlusions and perspective distortions inherent in single-view imaging result in large portions of the scene being partially invisible. This gap becomes especially challenging when trying to predict the structure and behavior of dynamic foreground objects, such as vehicles, pedestrians, and cyclists, which are critical for safe navigation in dynamic driving environments.

Monocular 3D Semantic Occupancy Prediction (SOP) [1]–[4] aims to infer a 3D semantic occupancy grid from a 2D image, effectively representing the geometry and semantics of the environment around the vehicle. Each spatial location in the grid records its occupancy and semantic information, enabling comprehensive environmental perception. However, the inherently ill-posed nature of this task poses significant challenges in diverse driving scenarios. Despite these difficulties, SOP has attracted significant attention for its potential to enhance the 3D perception capabilities of autonomous vehicles.

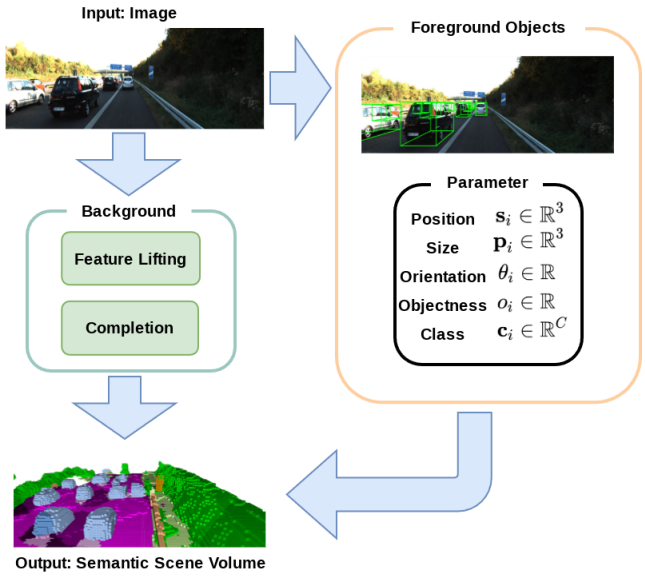

Fig. 1: OC-SOP predicts a semantic occupancy volume from a single image using a 2D-to-3D feature lifting and completion pipeline, while enhancing foreground object-centric awareness via explicit extraction of bounding box parameters. Parts of the output scene lie outside of the field of view (FoV), which are visualized as shadow areas.

Traditional semantic occupancy prediction methods largely extend the 2D semantic segmentation paradigm into the 3D domain. Approaches such as U-Net–based hierarchical feature extraction or Transformer-based self-attention, which capture voxel-to-voxel dependencies, primarily rely on local feature extraction and contextual reasoning. Traditional methods treat each category equally by merely adjusting category weights based on voxel counts without additional design, resulting in significant limitations. In autonomous driving scenarios, background categories (e.g., roads and buildings) typically yield better prediction results due to their simple shapes, consistent textures, and spatial continuity. In contrast, foreground objects (e.g., vehicles, pedestrians, and cyclists) are much more challenging to predict because they exhibit complex geometries and diverse appearances. Moreover, since these foreground objects actively participate in driving scenarios and display far more intricate behaviors and interactions, accurately predicting them is critical for safe driving.

This research has been supported by MBZIRC prize money. All authors are with the Autonomous Intelligent Systems group, Computer Science Institute VI – Intelligent Systems and Robotics – and the Center for Robotics and the Lamarr Institute for Machine Learning and Artificial Intelligence, University of Bonn, Germany; caoh@ais.uni-bonn.de

To address these challenges, we propose Object-Centric SOP (OC-SOP), a framework that incorporates high-level object-centric awareness into the SOP pipeline. As shown in Fig. 1, object parameters such as pose and size provide robust constraints on object perception, ensuring that the predicted boundaries more closely align with the actual object dimensions. As a result, OC-SOP significantly improves object prediction accuracy, reduces the misclassification of empty space as part of adjacent objects, and mitigates distortions caused by camera effects. These enhancements ultimately boost the system's overall environmental understanding and decision-making capabilities.

Our main contributions in this paper are:

- We introduce OC-SOP, a novel framework for 3D semantic occupancy prediction (SOP) that leverages high-level object-centric awareness by extracting object-centric cues using a detection-based branch and integrating these cues into the main completion branch.
- We propose a novel transformer module that tokenizes bounding box parameters into queries, which are then fused into the latent space of the completion U-Net.
- OC-SOP significantly improves the prediction accuracy for foreground objects and achieves state-of-the-art performance in camera-based semantic scene completion (SSC) on SemanticKITTI.

II. Related Work

A. Camera-based 3D Reconstruction

Camera-based 3D perception aligns with the natural human ability to form a holistic scene understanding from rich visual information. It is also increasingly favored in autonomous driving due to its cost-effectiveness. Early multi-view reconstruction [5]–[8] and SLAM [9], [10] studies estimated 3D geometry from corresponding 2D feature points using explicit mathematical constraints. However, their reliance on multi-view observations restricted agent movement and limited scene reconstruction quality. Eigen et al. [11] first introduced end-to-end monocular depth estimation using a CNN, enabling 3D reconstruction from a single 2D image. This addressed the ill-posed problem of 2D-to-3D reconstruction. Inspired by [11], some works focus on object-level single-view 3D reconstruction [12]–[15], typically employing an encoder-decoder structure to learn explicit or implicit geometry representations. Dahnert et al. [16] expands this to scene-level reconstruction by lifting 2D panoptic features into 3D space. However, since single-view reconstruction focuses only on visible regions, it exhibits weak performance in predicting occluded areas, significantly hindering holistic 3D scene understanding. To address this, scene completion is proposed to predict unseen geometry based on observable information.

B. Semantic Occupancy Prediction (SOP)

Although there is some debate within the community regarding the definition of Semantic Occupancy Prediction (SOP), the widely accepted view is that SOP aims to predict occupancy status and semantics within a grid-based scene volume, covering both visible and occluded areas [17]. Based on the input modality, SOP can be categorized into LiDAR-based SOP [18]–[20], Camera-based SOP [2], and Fusion-based SOP [21], among others. A broader task is Semantic Scene Completion (SSC), which jointly predicts geometry and semantics for both visible and occluded regions [22], [23]. However, scene representation is not limited to grid-based formats; it also includes explicit representations like point clouds [24] and meshes [25], as well as implicit representations such as Neural Radiance Fields (NeRF) [26]. Voxel-based representations are an effective way to model 3D scenes, as they discretize space into voxels, each storing attributes like occupancy, semantics, or learned feature vectors. Due to their compatibility with 3D convolutional operations, early LiDAR-based SOP [18], [19] methods commonly used U-Net architectures to predict occluded regions. Later, the 3D completion U-Net was extended to camera-based approaches. MonoScene [2] backprojects image features along optical rays to construct an initial voxel representation, which is then refined using a 3D completion U-Net. VoxFormer [27] introduced a Transformer-based approach, encoding spatial information as queries and context as key-value pairs, leveraging deformable attention for 3D-SOP.

Most SOP methods focus on scene-level occupancy, treating background and foreground objects similarly. However, foreground objects exhibit fine-grained shapes and complex dynamics, making them more challenging for models to learn. Our work creates object-centric representations through an object detection branch, enabling the network to better capture shape details and refine object category predictions.

C. Object-Centric Perception

Human perception is fundamentally object-centric, as our cognitive system naturally focuses on distinct objects, which forms the basis for advanced cognition and effective interaction with the world. Recent works have exploited this principle in both 3D object detection [28] and video prediction. For instance, inspired by the end-to-end paradigm of DETR [29], DETR3D [30] establishes learnable 3D object queries that link 2D images via camera projection matrices, while OCVP [31] decomposes video frames into object components and models their dynamics and interactions to generate future video frames. Leveraging object-centric awareness for semantic occupancy prediction is a novel direction. Unlike object detection—which outputs coarse bounding box representations—semantic occupancy prediction (SOP) produces finer, voxel-level semantic maps. Moreover, compared to video prediction, which typically focuses on visible regions, SOP must also predict occupancy in unseen or occluded areas, making the incorporation of object-centric awareness even more critical for accurate scene understanding.

III. Method

We consider a monocular 3D Semantic Occupancy Prediction (SOP) task that jointly predict the occupancy status

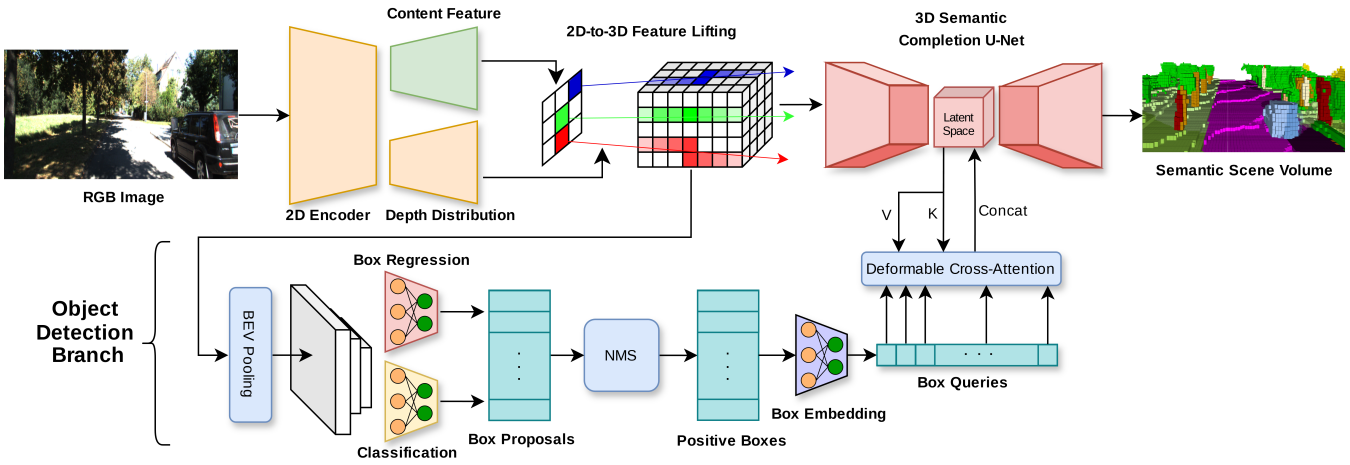

Fig. 2: The overall pipeline of OC-SOP can be divided into a main prediction branch (top) and an object detection-based branch (bottom). In the main branch, the RGB image is first fed into an encoder with dual decoders, where one decoder estimates depth and the other computes semantic content features. Next, a feature lifting module maps the content features into 3D space, and the resulting 3D features are processed by a 3D semantic completion U-Net that predicts the full semantic scene volume. In the object detection branch, the 3D semantic features are fed to a detection head to generate box proposals. Non-maximum suppression (NMS) is applied to select positive boxes. These positive boxes are then tokenized into queries via an MLP and fused into the completion network's latent space through deformable cross-attention.

and semantic label of every voxel in a scene. In this task, a single RGB image $\mathbf{X}^{\text{RGB}}$ is used as input to produce a 3D semantic volume $\hat{\mathbf{Y}}$ defined over a target space S with dimensions (H, W, D). Every voxel in $\hat{\mathbf{Y}}$ is assigned one of $M+1$ labels from the set $\mathcal{C} = \{c_0, c_1, \ldots, c_M\}$, where c_0 represents empty space and $c_1, \ldots, c_M$ correspond to the semantic classes.

As shown in Fig. 2, our dual-branch network architecture [32] comprises a main prediction branch and an object detection-based branch, which share a common feature lifting backbone. In the backbone, we adopt an encoder with dual decoders (EDD), where one decoder estimates depth and the other computes semantic content features. The semantic content features, combined with the estimated depth, are then fed into a feature lifting module that maps the semantic features into 3D space.

The resulting 3D semantic features are subsequently processed by both branches. The main branch employs a 3D semantic completion U-Net to predict the full semantic scene, while the object detection branch utilizes an object detection head to generate box proposals for the individual objects in the scene. These object detections are later fused into the latent space of the main branch, thereby enhancing the object-centric understanding.

A. Main Prediction Branch

Our main prediction network is composed of a feature lifting backbone and a 3D semantic completion U-Net, which represents the classic paradigm for vision-based semantic occupancy prediction originally proposed by Cao and de Charette in MonoScene [2]. MonoScene suffers from depth ambiguities caused by simply projecting 2D features along camera rays, which limits its ability to capture precise depth information. To address these issues, we enhance feature extraction by designing an encoder dual decoder (EDD).

Our encoder dual decoder (EDD) architecture disentangles semantic and geometric reasoning into two dedicated decoding branches. Starting from the shared 2D encoder feature map, one decoder branch extracts high-level semantic content, while the other estimates a dense per-pixel depth distribution over discretized depth bins. These two outputs are then combined through a depth-aware soft lifting process, where each 2D semantic feature is projected along its corresponding viewing ray, weighted by the predicted depth probabilities. This lifting operation is not limited to a single scale, but is performed hierarchically across multiple decoder layers, enabling the network to capture spatial correspondences at different levels of context.

Finally, the robust 3D features are fed into the 3D semantic completion U-Net, which predicts the complete semantic scene, including occluded regions, thereby significantly mitigating the camera effect. Our experiments reported in Sec. IV show that even without fusing object-centric awareness, our approach achieves improved performance compared to MonoScene.

B. Object Detection Branch

We feed the 3D semantic features also to an object detection network to detect individual objects in the scene. The detection network predicts a set of box proposals $\mathbf{B} \in \mathbb{R}^{N \times D} = \{\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_N\}$, where each proposal $\mathbf{b}_i$ is a D-dimensional vector representing the box parameters, formulated as $\mathbf{b}_i = [\mathbf{p}_i, \mathbf{s}_i, \sin\theta, \cos\theta, \mathbf{c}_i, o_i]$. Here, $\mathbf{p}_i \in \mathbb{R}^3$ denotes the center position, $\mathbf{s}_i \in \mathbb{R}^3$ represents the size, $\theta \in \mathbb{R}$ is the orientation (expressed via its sine and cosine), $\mathbf{c}_i \in \mathbb{R}^C$ indicates the class, and $o_i \in \mathbb{R}$ is the objectness score.

Inspired by center-based 3D object detection such as VoteNet [33] and PointRCNN [34], we adopt a lightweight detection head operating on lifted voxel features. The input

to the detection module is a 3D feature volume $\mathbf{F}^{3D} \in \mathbb{R}^{C' \times D' \times H' \times W'}$, where D' is the vertical dimension, and H' and W' correspond to the front-back and left-right axes in the bird's eye view (BEV), respectively.

We first perform average pooling along the vertical axis to obtain a BEV representation $\mathbf{F}^{\text{BEV}} \in \mathbb{R}^{C' \times H' \times W'}$, which is then processed by a shallow 2D CNN for spatial feature extraction. From the resulting BEV feature map, we generate a dense objectness heatmap $\mathbf{H} \in \mathbb{R}^{1 \times H' \times W'}$, indicating the likelihood of object centers at each spatial location. Based on the top-K peaks in this heatmap, we select candidate positions and apply two independent 2-layer MLPs at those locations: one for regressing the 3D bounding box parameters (center, size, and orientation), and the other for predicting objectness scores and semantic class probabilities.

For each candidate predicted from the heatmap, we assign a positive label if its center is within 1 m of any ground-truth object center, and a negative label if it is farther than 2 m, ensuring that proposals with different labels contribute appropriately to the respective loss computations.

The detection loss is given by:

$$L_{\text{det}} = L_{\text{obj}} + L_{\text{reg}} + L_{\text{cls}}, \tag{1}$$

where the objectness loss L_{obj} uses cross-entropy to optimize the model's ability to distinguish between positive and negative samples. Negative proposals, which do not have a corresponding ground truth box, are only used to compute the objectness loss, while their box parameters and class components are ignored. Proposals that are neither clearly positive nor negative do not contribute to the loss. For positive proposals, we compute both a regression loss L_{reg} and a semantic classification loss L_{cls}. The regression loss, implemented using Smooth L1, measures errors in the predicted box parameters (center position, size, and orientation), with different weights assigned to each component to balance their contributions. The semantic classification loss uses multi-class cross-entropy to ensure accurate category prediction.

C. Box Feature Fusion

After obtaining a set of box proposals, we apply non-maximum suppression (NMS) to select the final box features. In our approach, each selected box proposal is first encoded into a query vector using an MLP. For a given query q, we define K sampling points in the latent space, where each sampling location is determined by a base position p plus a learned offset Δp_k (for $k = 1, \ldots, K$). The deformable attention mechanism then computes the output as follows:

$$\text{Output}(q) = \sum_{k=1}^{K} A_k \, V\big(p + \Delta p_k\big), \tag{2}$$

where A_k denotes the attention weight at the kth sampling point and $V(\cdot)$ is the value function that extracts the corresponding feature from the latent space. This output is then fused into the latent space of the 3D semantic completion U-Net (via additional convolutional layers or residual fusion), thereby integrating complete query information with adaptively sampled latent features for effective fusion.

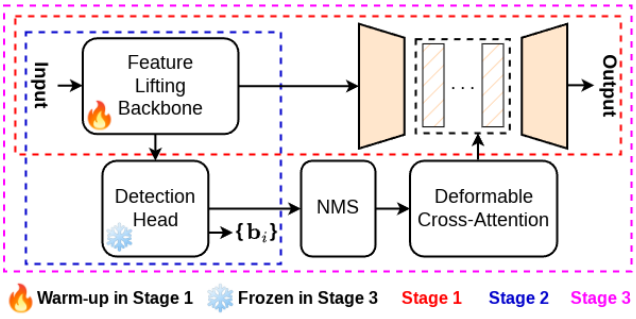

Fig. 3: Three-stage training strategy of OC-SOP.

D. Progressive Multi-task Training Strategy

Considering that our framework exhibits multi-task characteristics, in which the feature lifting backbone is utilized by both the main and detection branches, we adopt a three-stage training strategy, as shown in Fig. 3.

In Stage 1, we disable the detection branch and train only the feature lifting backbone along with the 3D completion U-Net. The goal in this phase is to warm up the backbone so that it efficiently extracts depth and semantic content information from the images and forms robust 3D features.

In Stage 2, we remove the 3D completion U-Net and load the pre-trained backbone to integrate the detection head module, with both positive and negative boxes used for loss computation.

In the final Stage 3, we freeze the detection head and apply non-maximum suppression (NMS) to select high-confidence, object-centric features. At this point, the overall network performance is further improved by effectively integrating the stable detection outputs into the fusion process.

This three-stage strategy is highly effective in multi-task learning scenarios for stabilizing training and achieving the reported results.

IV. Experiment

A. Experiment Setup

1) Datasets: We train our model using two datasets. For voxel-wise prediction, we employ the SemanticKITTI [35] dataset, on which numerous classic baselines have been validated. Built on the KITTI odometry benchmark, SemanticKITTI enables to study semantic segmentation through manual annotations on raw LiDAR point clouds. The semantic scene completion benchmark is constructed by accumulating annotated scans within sequences and voxelizing them to form a semantic scene volume with dimensions of 256×256×32 (with each voxel representing 0.2 m). The dataset contains 21 classes, including 19 semantic labels, one free label, and one unknown label. In our experiments, we use the RGB images from cam-2 with a resolution of 1226×370 (cropped to 1220×370) and follow the official training/validation split of 3834 and 815 samples, respectively.

TABLE I: Quantitative results on the SemanticKITTI test set.

Method	IoU	mIoU	Objects								Background										
			car (3.92%)	**truck** (0.16%)	**bicycle** (0.03%)	**motorcycle** (0.03%)	**other-vehicle** (0.20%)	**person** (0.07%)	**bicyclist** (0.07%)	**motorcyclist** (0.05%)	**road** (15.30%)	**parking** (1.12%)	**sidewalk** (11.13%)	**other-ground** (0.56%)	**building** (14.10%)	**fence** (3.90%)	**vegetation** (39.30%)	**trunk** (0.51%)	**terrain** (9.17%)	**pole** (0.29%)	**traffic-sign** (0.08%)
LMSCNet* [18]	31.38	7.07	14.30	0.30	0.00	0.00	0.00	0.00	0.00	0.00	46.70	13.50	19.50	3.10	10.30	5.40	10.80	0.00	10.40	0.00	0.00
AICNet* [1]	23.93	7.09	15.30	0.70	0.00	0.00	0.00	0.00	0.00	0.00	39.30	19.80	18.30	1.60	9.60	5.00	9.60	1.90	13.50	0.10	0.00
JS3C-Net* [19]	34.00	8.97	20.10	0.80	0.00	0.00	4.10	0.00	0.20	0.20	47.30	19.90	21.70	2.80	12.70	8.70	14.20	3.10	12.40	1.90	0.30
MonoScene [2]	34.16	11.08	18.80	3.30	0.50	0.70	**4.40**	1.00	1.40	0.40	54.70	24.80	27.10	5.70	14.40	11.10	14.90	2.40	19.50	3.30	2.10
TVPFormer [4]	34.25	11.26	19.20	3.70	1.00	0.50	2.30	1.10	2.40	0.30	55.10	**27.40**	27.20	6.50	14.80	11.00	13.90	2.60	20.40	2.90	1.50
VoxFormer [27]	42.95	12.20	20.80	3.50	1.00	0.70	3.70	1.40	2.60	0.20	53.90	21.10	25.30	5.60	19.80	11.10	**22.40**	7.50	21.30	5.10	4.90
OccFormer [3]	34.53	12.32	21.60	1.20	1.50	1.70	3.20	2.20	1.10	0.20	55.90	31.50	**30.30**	6.50	15.70	11.90	16.80	3.90	21.30	3.80	3.70
OC-SOP	**43.30**	**14.83**	**30.40**	**4.80**	**7.80**	**6.30**	3.50	**5.80**	**3.70**	**1.30**	**56.00**	27.10	29.50	**7.10**	**21.50**	**13.30**	20.80	**8.30**	**22.20**	**5.90**	**6.40**

All listed baselines are vision-based methods. * denotes the reproduced results reported by [2].
IoU focuses solely on the occupancy status while mIoU evaluates individual semantic categories. **Best** and second best results are highlighted.

For detection branch training, we utilize the KITTI [36] 3D object detection benchmark dataset, which comprises 7481 training images and 7518 test images along with their corresponding point clouds. We follow the train/validation split defined by Chen et al. [37], dividing the training set into 3712 training images and 3769 validation images. KITTI annotates three foreground object categories (Car, Pedestrian, and Cyclist), which completely align with our definition of foreground objects and are well suited for object-centric tasks on SemanticKITTI. Although the KITTI's semantic labels are coarser than the fine-grained semantic labels in SemanticKITTI, these rough labels still provide a semantic prior for foreground objects in semantic occupancy prediction tasks.

2) Implementation Details: The model is trained on an NVIDIA A6000 GPU in three stages: a warm-up Stage 1 for 5 epochs, a detection branch training Stage 2 for 10 epochs, and a joint fine-tuning Stage 3 for 10 epochs. During joint fine-tuning and inference, we set the detection head's objectness threshold to 0.2 to filter out low-score negative predictions, followed by applying an NMS module to obtain the final box proposals. The NMS IoU threshold is set to 0.7.

B. Main Results

We evaluate our model on the SemanticKITTI test set using two metrics. Intersection over Union (IoU) metrics focus only on occupancy status without considering semantic labels. Mean IoU (mIoU) is used to evaluate the accuracy of the semantic labels. Specifically, the IoU for each semantic category is first computed, and then the average of these values is taken to reflect overall performance. All IoU calculations exclude the unknown category. The results are reported in Tab. I, where we also compare our approach with several vision-based baselines on the SemanticKITTI benchmark. The results show that OC-SOP significantly outperforms all baselines on foreground object categories. For instance, when considering only foreground object categories, OC-SOP achieves an mIoU_{obj} of 7.95%, which is 3.71 percentage points higher than the second-best method, VoxFormer, which achieves an mIoU_{obj} of 4.24%. Moreover, even though background categories do not directly benefit from object-centric awareness, the design of our EDD and 3D semantic completion U-Net still achieves comparable performance, slightly outperforming most baselines.

Fig. 4 shows several qualitative results from the SemanticKITTI validation set. We compare our predictions with those of MonoScene to highlight the benefits of incorporating object-centric awareness. In Scene A, multiple vehicles are parked along both sides of the road with narrow gaps between them. MonoScene merges these vehicles into one block, making it difficult to identify the number of vehicles and resulting in severely distorted shapes. In contrast, our method clearly distinguishes each vehicle, preserving distinct gaps and maintaining shape integrity. In Scene B, a bicyclist appears in front of the ego-vehicle. MonoScene misclassifies the bicyclist as a motorcyclist, and its shape is significantly distorted by camera effects. Our predictions, guided by bounding box constraints, more accurately capture the bicyclist's shape. It is important to note that since SemanticKITTI's ground truth is derived from aggregating multiple frames, dynamic objects may exhibit ghosting artifacts, a limitation that partly contributes to the misclassifications observed in MonoScene. Scene C presents a similar issue with a bicyclist, and in addition, MonoScene fails to detect vehicles near the field-of-view boundaries. This suggests that traditional SOP methods are less sensitive to objects that are severely truncated in the view frustum, even though such objects, often located near the ego-vehicle, are vital for accurate scene understanding. Finally, in Scene D, MonoScene mistakenly predicts a vehicle in an area where none exists, likely due to misleading lighting and shadow cues. The error from the semantic segmentation module is not corrected by the reprojection process and thus appears in the final output. In practical applications, this could lead to incorrect estimation of dynamic variables such as speed

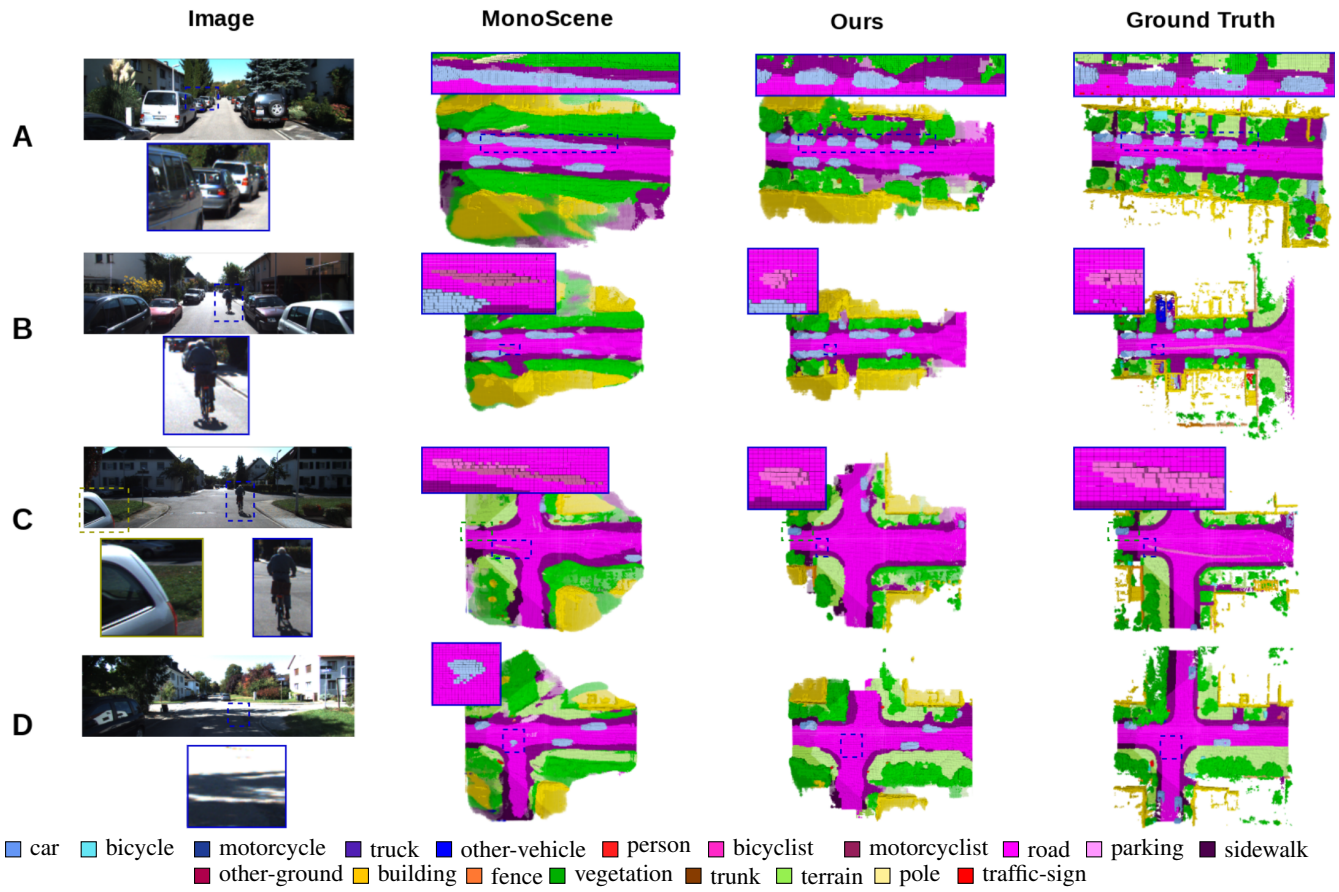

Fig. 4: Qualitative results on the SemanticKITTI validation set. Only the 19 semantic classes are visualized; voxels labeled as empty space are omitted for clarity. Voxels located in unknown regions are rendered with 20% opacity to reflect uncertainty. Parts of the scene that lie outside the field of view (FoV) are visualized as shaded areas.

and trajectory, thereby affecting the comfort and even the safety of passengers and other traffic participants. In our framework, however, the objectness score helps to correct such misestimations, preventing these hallucinations in the final prediction.

C. Ablation Studies

We conducted ablation studies on the SemanticKITTI validation set to assess the contributions of each sub-module.

In Setting I, we employ the full model, which achieves the best performance for all foreground object categories and most background categories, yielding overall best results.

In Setting II, to isolate the impact of the box feature fusion module, we replace the deformable cross attention with spatial concatenation fusion. Specifically, each box proposal's feature is first transformed using an MLP and then concatenated along the channel dimension with the main branch's latent space. Subsequent convolution layers fuse these features. The results indicate that although Setting II still incorporates object-centric awareness, the deformable cross attention used in Setting I more effectively integrates box proposals with the latent features. This leads to a notable improvement in foreground object performance. For background categories, the performance of Setting II is slightly lower than that of Setting I, but the difference is not as pronounced as for foreground objects.

In Setting III, only the main prediction branch is used for inference without any object-centric awareness. Comparing Setting I with Setting III, we observe that incorporating object-centric awareness significantly enhances the estimation of foreground objects, with performance improving by approximately 4.54 percentage points (from 3.41% to 7.95%) while background categories show only a modest improvement of around 1.02 percentage points (from 18.81% to 19.83%). Notably, the main prediction branch consists of an encoder dual decoder (EDD) and a 3D completion U-Net. This architecture effectively enhances the extraction of depth information, thereby mitigating camera effects, and still achieves competitive performance that slightly outperforms MonoScene.

Finally, Setting IV retains only the 3D semantic completion U-Net, which relies solely on the raw image and an external depth predictor to form a pseudo-LiDAR 3D map. Due to the lack of a robust feature extractor, Setting IV shows poor performance, with limited ability to predict several foreground categories and weak performance on background categories. The comparison between Settings III and IV clearly demonstrates the effectiveness of the EDD in enhancing feature extraction for semantic occupancy prediction.

TABLE II: Ablation study of each sub-module on the SemanticKITTI validation set.

Setting		I	II	III	IV
Dual Decoder		✓	✓	✓	-
Detection Branch		✓	✓	-	-
Deformable Cross-attention		✓	-	-	-
Objects	car	**30.40**	25.60	17.80	17.30
	truck	**4.80**	3.90	3.10	0.70
	bicycle	**7.80**	5.30	0.40	0.00
	motorcycle	**6.30**	4.20	0.70	0.00
	other-vehicle	**3.50**	3.40	2.50	0.00
	person	**5.80**	4.40	1.50	0.00
	bicyclist	**3.70**	3.10	1.10	0.00
	motorcyclist	**1.30**	1.20	0.20	0.00
Background	road	**56.00**	55.80	55.30	46.30
	parking	27.10	26.50	**27.70**	12.10
	sidewalk	**29.50**	29.10	28.20	20.10
	other-ground	7.10	6.90	**7.30**	4.20
	building	**21.50**	19.90	18.70	10.70
	fence	**13.30**	12.90	10.90	6.90
	vegetation	**20.80**	20.50	19.00	10.30
	trunk	8.30	**8.50**	6.30	1.30
	terrain	22.20	21.50	**22.90**	11.20
	pole	**5.90**	5.50	4.80	0.30
	traffic-sign	**6.40**	6.10	5.80	0.10
IoU		**43.30**	41.98	41.52	29.78
mIoU		**14.56**	13.92	12.33	7.45

IoU focuses solely on the occupancy status while mIoU evaluates individual semantic categories. **Best** and second best results are highlighted.

V. Conclusions and Outlook

We propose OC-SOP, which innovatively integrates object-centric awareness into the semantic occupancy prediction task. Our results on the SemanticKITTI test set demonstrate that OC-SOP significantly improves the accuracy of foreground object predictions. By focusing on other traffic participants, our approach enhances the perception system's ability to understand complex traffic scenarios, ultimately contributing to safer decision-making in autonomous driving. In future work, we plan to explore emerging techniques in prompt-driven representation learning [38]–[40] and multi-modal fusion [41], [42] to further improve object-centric semantics and generalization across diverse sensing conditions.

References

[1] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3D semantic scene completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3351–3359.

[2] A.-Q. Cao and R. de Charette, "MonoScene: Monocular 3D semantic scene completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3991–4001.

[3] Y. Zhang, Z. Zhu, and D. Du, "OccFormer: Dual-path transformer for vision-based 3d semantic occupancy prediction," *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[4] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3D semantic occupancy prediction," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[6] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 8, pp. 1362–1376, 2009.

[7] Y. Hu, J. Zhang, Z. Zhang, R. Weilharter, Y. Rao, K. Chen, R. Yuan, and F. Fraundorfer, "ICG-MVSNet: Learning intra-view and cross-view relationships for guidance in multi-view stereo," *arXiv preprint arXiv:2503.21525*, 2025.

[8] Y. Hu, T. Fu, G. Niu, Z. Liu, and M.-O. Pun, "3D map reconstruction using a monocular camera for smart cities," *Journal of Supercomputing*, 2022.

[9] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 834–849.

[10] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[12] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[13] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 605–613.

[14] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4460–4470.

[15] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 523–540.

[16] M. Dahnert, J. Hou, M. Nießner, and A. Dai, "Panoptic 3D scene reconstruction from a single RGB image," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8282–8293, 2021.

[17] S. Su, N. Chen, F. Juefei-Xu, C. Feng, and F. Miao, "α-OCC: Uncertainty-aware camera-based 3D semantic scene completion," *arXiv preprint arXiv:2406.11021*, 2024.

[18] L. Roldao, R. de Charette, and A. Verroust-Blondet, "LMSCNet: Lightweight multiscale 3D semantic completion," in *International Conference on 3D Vision (3DV)*, 2020, pp. 111–119.

[19] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion," in *National Conference on Artificial Intelligence (AAAI)*, 2021, pp. 3101–3109.

[20] H. Cao, R. Materla, and S. Behnke, "SWA-SOP: Spatially-aware window attention for semantic occupancy prediction in autonomous driving," *arXiv preprint arXiv:2506.18785*, 2025.

[21] H. Cao and S. Behnke, "SLCF-Net: Sequential LiDAR-camera fusion for semantic scene completion using a 3D recurrent U-Net," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 2767–2773.

[22] L. Roldão, R. de Charette, and A. Verroust-Blondet, "3D semantic scene completion: A survey," *International Journal of Computer Vision (IJCV)*, vol. 130, no. 8, pp. 1978–2005, 2022.

[23] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1746–1754.

[24] H. Cao and S. Behnke, "DiffSSC: Semantic LiDAR scan completion using denoising diffusion probabilistic models," *arXiv preprint arXiv:2409.18092*, 2024.

[25] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner, "ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4578–4587.

[26] T.-A.-Q. Nguyen, A. Bourki, M. Macudzinski, A. Brunel, and M. Bennamoun, "Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review," *arXiv preprint arXiv:2402.11141*, 2024.

[27] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9087–9098.

[28] M. Wang, D. Li, J. R. Casas, and J. Ruiz-Hidalgo, "Adaptive fusion of lidar features for 3d object detection in autonomous driving," *Sensors*, 2025.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[30] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proceedings of Machine Learning Research (PMLR)*, 2022, pp. 180–191.

[31] A. Villar-Corrales, I. Wahdan, and S. Behnke, "Object-centric video prediction via decoupling of object dynamics and interactions," in *IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 1234–1238.

[32] H. Tao, J. Li, Z. Hua, and F. Zhang, "DUDB: Deep unfolding-based Dual-Branch feature fusion network for pan-sharpening remote sensing images," 2023.

[33] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Deep Hough voting for 3D object detection in point clouds," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[34] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[37] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2147–2156.

[38] L. Zhou, N. Li, M. Ye, X. Zhu, and S. Tang, "Source-free domain adaptation with class prototype discovery," *Pattern Recognition*, 2024.

[39] N. Li, M. Ye, L. Zhou, S. Tang, Y. Gan, Z. Liang, and X. Zhu, "Self-Prompting analogical reasoning for UAV object detection," in *National Conference on Artificial Intelligence (AAAI)*, 2025.

[40] W. Wu, X. Qiu, S. Song, Z. Chen, X. Huang, F. Ma, and J. Xiao, "Prompt categories cluster for weakly supervised semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 3198–3207.

[41] W. Wu, Z. Chen, X. Qiu, S. Song, X. Huang, F. Ma, and J. Xiao, "LLM-Enhanced multimodal fusion for Cross-Domain sequential recommendation," *arXiv preprint arXiv:2506.17966*, 2025.

[42] W. Wu, S. Song, X. Qiu, X. Huang, F. Ma, and J. Xiao, "Image fusion for cross-domain sequential recommendation," 2025.