Enhanced Geometry and Semantics for Camera-Based 3D Semantic Scene Completion

Haihong Xiao, Wenxiong Kang, *Member, IEEE*, Yulan Guo, *Senior Member, IEEE*, Hao Liu, and Ying He, *Member, IEEE*

***Abstract*—Giving machines the ability to infer the complete 3D geometry and semantics of complex scenes is crucial for many downstream tasks, such as decision-making and planning. Vision-centric Semantic Scene Completion (SSC) has emerged as a trendy 3D perception paradigm due to its compatibility with task properties, low cost, and rich visual cues. Despite impressive results, current approaches inevitably suffer from problems such as depth errors or depth ambiguities during the 2D-to-3D transformation process. To overcome these limitations, in this paper, we first introduce an Optical Flow-Guided (OFG) DepthNet that leverages the strengths of pretrained depth estimation models, while incorporating optical flow images to improve depth prediction accuracy in regions with significant depth changes. Then, we propose a depth ambiguity-mitigated feature lifting strategy that implements deformable cross-attention in 3D pixel space to avoid depth ambiguities caused by the projection process from 3D to 2D and further enhances the effectiveness of feature updating through the utilization of prior mask indices. Moreover, we customize two subnetworks: a residual voxel network and a sparse UNet, to enhance the network's geometric prediction capabilities and ensure consistent semantic reasoning across varying scales. By doing so, our method achieves performance improvements over state-of-the-art methods on the SemanticKITTI, SSCBench-KITTI-360 and Occ3D-nuScene benchmarks.**

***Index Terms*—3D vision, semantic scene completion, SemanticKITTI benchmark.**

Received 16 February 2025; revised 13 September 2025; accepted 10 November 2025. Date of publication 24 December 2025; date of current version 16 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62376100, Grant 62076086, and Grant 62476077; and in part by Nanyang Technological University through China Scholarship Council (CSC). The associate editor coordinating the review of this article and approving it for publication was Dr. Xinfeng Zhang. *(Corresponding author: Wenxiong Kang.)*

Haihong Xiao is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230088, China (e-mail: hhxiao@hfut.edu.cn).

Wenxiong Kang is with the School of Automation Science and Engineering and the School of Future Technology, South China University of Technology, Guangzhou 510641, China, and also with Pazhou Laboratory, Guangzhou 510335, China (e-mail: auwxkang@scut.edu.cn).

Yulan Guo is with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen Campus, Shenzhen 510006, China (e-mail: guoyulan@sysu.edu.cn).

Hao Liu is with the School of Geospatial Artificial Intelligence, the Key Laboratory of Geographic Information Science (Ministry of Education), and the Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China (e-mail: hliu@geoai.ecnu.edu.cn).

Ying He is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: yhe@ntu.edu.sg).

Digital Object Identifier 10.1109/TIP.2025.3635475

I. Introduction

SEMANTIC Scene Completion (SSC) aims to reconstruct voxel-wise semantics of a 3D scene from partial LiDAR point clouds or limited image inputs, which has become an emerging and pivotal task in autonomous driving as it provides a more fine-grained description of the 3D world compared to conventional 3D object detection.

Recent progress in SSC can be broadly divided into LiDAR-based methods and camera-based methods, depending on the input. Although LiDAR-based methods have made remarkable progress due to their accurate depth information, LiDAR sensors are expensive and LiDAR sweeps are typically sparse and incomplete, limiting their widespread application in practice. Recent innovations driven by BEVFormer [1] and its variants [2], [3] have enabled precise predictions of geometry and semantics by using only images due to their favorable color distinguishability, providing a cost-effective alternative. This has also further spawned a related field: semantic occupancy prediction [4]. Although previous studies [5], [6] sometimes equate semantic occupancy prediction with SSC, subtle definitional differences remain. Semantic occupancy prediction primarily focuses on using occupancy grids to represent the entire 3D scene [7], typically with surround-view inputs. However, SSC emphasizes completing scenes within the camera's field of view from limited observations, such as single-view images, stereo images, or partial point clouds. Considering the computational demands and ease of deployment, this paper primarily focuses on the camera-based SSC task.

MonoScene [8], as a pioneering work, lifts 2D features into 3D space using the Feature Line of Sight Projection (FLoSP) module. However, due to its depth-agnostic flaw, this lifting operation occasionally maps 2D features into incorrect spatial areas, such as empty or overlapping semantic regions. Inspired by this, subsequent studies have expanded the SSC community along two new avenues: lift-splat-based [9], [10], [11] and 3D-to-2D projection [2], [3], [12] methods. Lift-splat-based methods mainly utilize the Lift-Splat-Shot (LSS) paradigm [11] to transform image features into pseudo LiDAR features by executing the outer product between context features and discrete depth distributions. 3D-to-2D projection methods project predefined voxel queries onto the image plane using camera intrinsic and extrinsic parameters and then apply 2D deformable attention [13] to aggregate relevant image features, thus updating the voxel queries. Despite their impressive results, these methods have inherent limitations:

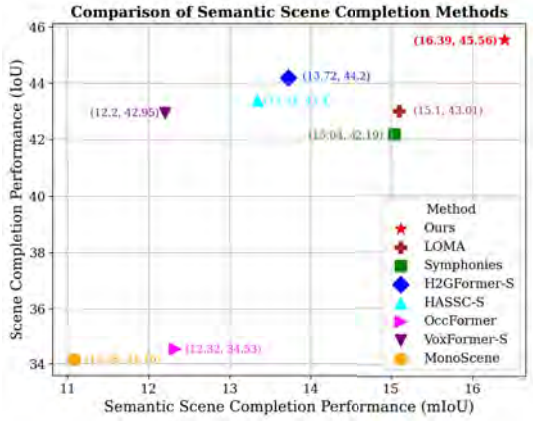

Fig. 1. Comparison with state-of-the-art camera-based SSC methods on the SemanticKITTI test set.

lift-splat-based methods struggle with depth errors due to their single-step process in feature lifting, while 3D-to-2D projection methods may result in entangled aggregated features due to depth ambiguities during projection.

Owing to the astounding success of 3D Deformable Attention (DFA3D) [21] in 3D object detection, we employ DFA3D as our foundational component in this paper. Meanwhile, we leverage prior mask indices generated from the pretrained depth estimation model to guide effective feature updating in DFA3D, thereby achieving depth ambiguity-mitigated feature lifting. Notably, the concurrent work CGFormer [19] also employs DFA3D in the 2D-to-3D transformation process. However, it fails to address the consistent reasoning of the differences between objects and backgrounds in the scene, as well as the inaccuracies in the depth prediction for moving objects. We optimize our approach in two aspects: *Depth Estimation Refinement* and *Geometric-Semantic Enhancement*. *Depth Estimation Refinement*: unlike the original DFA3D, which directly uses the Convolutional Neural Network(CNN) [22] in DepthNet for depth estimation, we harness the advantages of pretrained depth estimators and further introduce dual fusion cross-attention to refine depth distribution. However, due to insufficient geometric constraints, pretrained depth estimation models often result in depth inaccuracies, particularly in areas with significant depth changes (depth jumps). To address this, we present an optical flow modulation module to optimize the depth maps. The key insight is that regions with significant optical flow changes typically correspond to those with significant depth changes, such as car boundaries. *Geometric and Semantic Enhancement*: we carefully analyze the task characteristics and find that introducing a residual voxel network after deformable self-attention could further enhance the network's geometric and detail prediction capabilities. Additionally, given the high performance of UNet [23] in segmentation tasks, we replace the conventional 3D convolution [12] or deconvolution [8] in the segmentation head with a more efficient sparse UNet structure, which adopts an Adaptive Hierarchical Aggregator (AHA) [24] to achieve consistent semantic reasoning across varying scales, such as car and road.

By doing so, our method demonstrates its superiority on the SemanticKITTI test set, as demonstrated in Fig. 1. We will release our code and trained models. To clearly illustrate the distinctions between our method and existing approaches, we also present a comprehensive comparison in Tab. I, which highlights key aspects such as depth ambiguity-mitigated feature lifting, motion-aided object perception, scale-aware semantic perception, fine-grained detail generation, as well as the backbone and 2D-3D lifting techniques utilized by each method. The contributions of this work can be summarized as follows:

- We present an Optical Flow-Guided (OFG) DepthNet to improve depth prediction accuracy in regions with significant depth changes.
- We propose a depth ambiguity-mitigated feature lifting strategy and further enhance feature updating effectiveness by introducing prior mask indices.
- We tailor two subnetworks: a residual voxel network and a sparse UNet, which further enhance the network's geometric prediction capabilities and achieve consistent semantic reasoning across varying scales.
- We evaluate our method on the SemanticKITTI, SSCBench-KITTI-360 and Occ3D-nuScene public benchmarks. It achieves state-of-the-art semantic completion quality in terms of IoU and mIoU.

II. Related Work

A. View Transformation

Advances in camera-based 3D perception have allowed for transforming 2D features into corresponding 3D space using only image input. These view transformation techniques can be broadly categorized into three types: 1) lift-splat-based methods [9], [10], [11], [25], 2) 3D-to-2D projection methods [2], [12], [17], and 3) 2D-to-3D network transformation methods [26], [27]. Lift-splat-based methods utilize the CNN to extract contextual features from input images while estimating discrete depth distributions. The contextual features are then lifted into 3D space by combining them with depth information through an outer product operation. 3D-to-2D projection methods primarily project predefined voxel queries onto the image plane using camera intrinsic and extrinsic parameters and then employ 2D deformable attention [13] to aggregate relevant image features. To efficiently achieve 3D-to-2D interaction, these methods often leverage pretrained depth estimators [28] to generate binary occupancy query proposals. 2D-to-3D network transformation methods implicitly construct view transformation relationships using neural networks, with architectures including Multi-Layer Perceptron (MLP) and Transformer. Despite progress, these methods exhibit significant limitations: they suffer from depth estimation inaccuracies near moving object edges, are prone to feature entanglement due to depth ambiguities, and struggle to model spatial relationships in complex outdoor scenes.

B. Scene Completion

Scene completion aims to fill missing regions in 3D space by generating plausible geometric structures and realistic textures. Scene representations can be broadly categorized into

two types: explicit [29], [30], [31], [32], [33] and implicit [34], [35] representations.

Classical explicit representations, such as point clouds [29], meshes [31], and voxels [30], dominated early research; as a result, scene completion primarily focused on geometry completion. Two primary strategies are commonly adopted: 1) traditional methods based on object decomposition and replacement [36], [37], and 2) learning-based scene generation methods [38], [39], [40], [41]. However, traditional methods rely heavily on handcrafted priors, which limit their adaptability. In contrast, learning-based methods, while more flexible, often encounter challenges such as data sparsity and computational inefficiency when applied to complex scenes. Recently, implicit scene representations, such as Neural Radiance Fields (NeRF) [34] and 3D Gaussian Splatting (3DGS) [35], have achieved significant progress. NeRF represents scenes as continuous functions by modeling geometry and appearance using parameterized Multi-Layer Perceptrons (MLPs) and compactly posed images via differentiable volume rendering [42]. Some works incorporate textual priors (e.g., CLIP [43]) or semantic features (e.g., DINO [44]) to enhance 3D semantic understanding, thereby improving completion quality. There are also some studies that combine 2D image inpainting models with segmentation or depth priors [45], [46] to jointly optimize NeRF. However, NeRF's inherently slow sampling speed remains a limitation for practical applications. Inspired by recent advancements in 3DGS, some studies [47] have begun exploring scene completion methods based on 3DGS. Compared to NeRF, 3DGS offers significant advantages in rendering efficiency and completion quality.

C. Semantic Scene Completion

Previous studies primarily focus on indoor scenes, utilizing depth maps [48] or additional RGB images [49] as inputs. Significant progress has been made in outdoor scenes with models trained on larger datasets like SemanticKITTI [50], which typically falls into one of two categories: LiDAR-based [51], [52], [53], [54], [55] and camera-based [2], [4], [7], [8], [9], [17], [56], [57] approaches.

Extensions of LiDAR-based methods study multi-scale supervised learning [51], point-voxel interaction [52], knowledge distillation [53], and local implicit functions [54]. Recently, Cao et al. [55] proposed a panoptic scene completion framework that further enhances attention to instance objects and uncertainty estimation. However, the high cost of LiDAR sensors, as well as the sparsity and varying density of point clouds, limit their application. Camera-based approaches [2], [8], [9], [17] are becoming a promising path due to their low cost of deployment. This has also further promoted the development of surround-view semantic occupancy prediction [4], [7], [56]. These methods either utilize pretrained depth estimators to generate query proposals [2] and then adopt 2D deformable attention to update projected features, or directly estimate depth distributions [9] to lift 2D features into 3D space, or incorporate multi-frame images (historical frames) [5] to enhance the completeness of 3D scenes. Albeit effective, projection-based methods often lead to entangled features due to depth ambiguity and direct depth estimation cannot effectively handle depth errors owing to its single-step nature. Meanwhile, they have not effectively addressed the issue of dynamic object changes during the multi-frame fusion process. Additionally, several notable works have attempted to enhance scene understanding by incorporating language knowledge [58] or segmentation priors [59], providing valuable insights for future research. Considering the label-intensive annotations required in SSC, recent methods also adopt self-supervised ways based on NeRF [60] and Gaussian splatting [61]. However, these typically involve trade-offs between time and quality and are, hence, not directly applicable to the autonomous driving task, which demands high security and adaptability to rapidly changing scenarios.

III. Method

A. Problem Setup

Given RGB images as input, the camera-based semantic scene completion aims to predict voxel-wise semantics within a specific visual range in front of the car. Specifically, we use the images $\{I_{left}, I_{right}\}$ captured by the left and right cameras as input. The model predicts the semantic voxel grid $V \in \{v_0, v_1, \ldots, v_c\}^{H \times W \times Z}$, where each voxel is either occupied or empty, represented by different semantic classes or as empty. Here, c denotes the total number of semantic classes, and H, W, Z represent the length, width, and height of the voxel grid, respectively.

The overall objective is to train a neural model Φ to predict the semantic voxel labels, making them as close as possible to the ground truth. It is important to note that some previous methods, which define monocular semantic scene completion, do not strictly rely on a single image. This is because they overlook the need for stereo image input in depth estimation models. In this paper, we define the input as stereo images to clarify this point.

B. Overall Architecture

This section presents a semantic scene completion framework that enhances both geometry and semantics. As shown in Fig. 2, the proposed method takes stereo images as input and generates semantic scene information for the current viewpoint. First, the images are processed using offline depth estimation and optical flow models to compute the corresponding depth and optical flow maps, which assist in estimating the scene's geometric information, as demonstrated by OFG DepthNet. Next, unlike previous methods [2], [17] that predict coarse-grained occupancy queries and enhance them through interactions with image features, this paper directly performs information interaction and enhancement in 3D space, thereby avoiding depth ambiguity issues that may arise from query projection. To achieve this, we propose a depth ambiguity-mitigated feature lifting strategy. Then, considering the scale differences between the background and various objects, and further enhancing the geometric representation, we design two subnetworks for semantic and geometric enhancement. Finally, we introduce the loss function used for training.

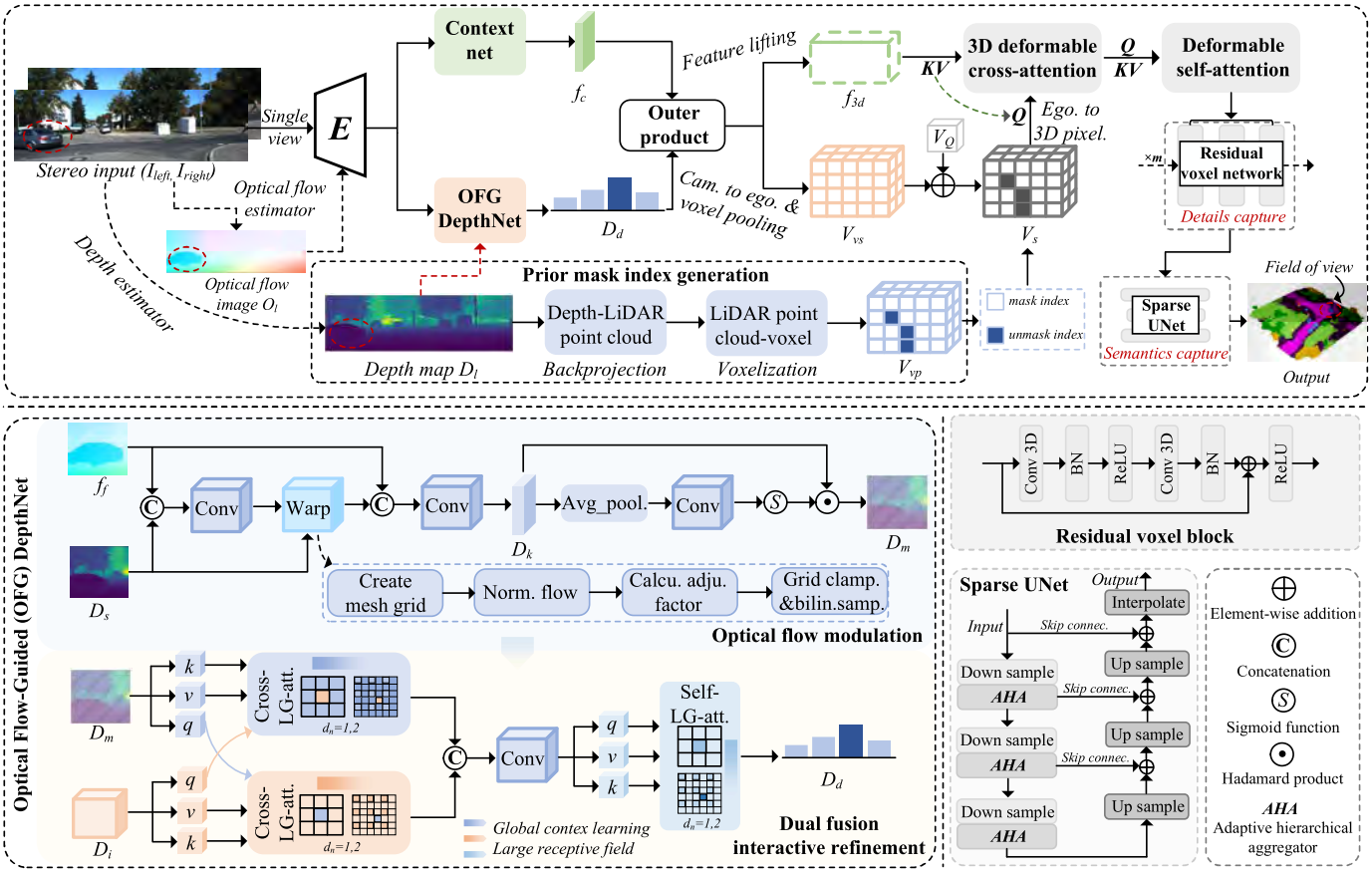

Fig. 2. Overall framework of our proposed method. First, we enhance depth prediction accuracy in OFG DepthNet by leveraging depth maps and optical flow images estimated by offline estimators. Subsequently, we employ 3D deformable cross-attention to facilitate effective feature interactions in 3D pixel space. In addition, we further use prior mask indices to guide initial voxel query updating. Finally, we utilize a residual voxel network and a sparse UNet to capture geometric details and ensure consistent semantic reasoning across varying scales.

TABLE I

COMPARISON AMONG SOME REPRESENTATIVE METHODS. "FLOSP" REPRESENTS THE FEATURES LINE OF SIGHT PROJECTION. "TPV" REFERS TO THE TRI-PERSPECTIVE VIEW. "BEV" REPRESENTS THE BIRD'S EYE VIEW. RESNET-50-MD INDICATES RESNET50 WITH WEIGHTS INITIALIZED BY MASKDINO. EFFICIENTNETB7-FPN INDICATES THAT EFFICIENTNETB7 IS AUGMENTED WITH FPN. "DFA3D" REPRESENTS THE 3D DEFORMABLE ATTENTION

Method	depth ambiguity-mitigated feature lifting	motion-aided object perception	scale-aware semantic perception	fine-grained detail generation	image backbone	2D-3D lifting
MonoScene [CVPR22] [8]	✗	✗	✗	✗	EfficientNetB7	FLoSP-based
TPVFormer [CVPR23] [7]	✗	✗	✗	✗	EfficientNetB7	View-TPV
VoxFormer-S [CVPR23] [2]	✗	✗	✗	✗	ResNet-50	3D-2D projection
StereoScene [IJCAI24] [14]	✗	✗	✗	✗	EfficientNetB7	View-BEV& stereo
H2GFormer-S† [AAAI24] [3]	✗	✗	✗	✗	ResNet-50	3D-2D projection
SparseOcc [CVPR24] [15]	✗	✗	✗	✗	EfficientNetB7-FPN	Lift-splat-based
HASSC-S† [CVPR24] [16]	✗	✗	✗	✗	ResNet-50	3D-2D projection
Symphonies [CVPR24] [17]	✗	✗	✗	✗	ResNet-50-MD	3D-2D projection
OccGen [ECCV24] [18]	✓	✗	✗	✓	EfficientNetB7-FPN	DFA3D
CGFormer [NeurIPS24] [19]	✓	✗	✗	✓	EfficientNetB7	DFA3D
LOMA [AAAI25] [20]	✗	✗	✗	✓	ResNet-50	3D-2D projection
Ours	✓	✓	✓	✓	EfficientNetB7	DFA3D

C. Optical Flow-Guided DepthNet

To fully leverage the advantages of pretrained depth estimation models and effectively integrate them with DepthNet, while addressing the depth jumps in depth maps caused by insufficient geometric constraints, we introduce an Optical Flow-Guided (OFG) DepthNet.

We use images I_{left} and I_{right} as inputs and adopt the pretrained models MobileStereoNet [62] and GMFlow [63] to obtain the depth map and optical flow image of I_{left}, denoted as D_l and O_l, respectively. Next, we use the pretrained EfficientNet-B7 [64] model to extract 2D features from I_{left} and O_l, denoted as f_r and f_o, respectively. Then, we employ a custom context network and an OFG DepthNet to obtain the corresponding context features f_c and depth estimation distributions D_d. In the context network, following [19], we utilize the MLP, Squeeze-and-Excitation (SE) layer and 2D CNN to process camera parameters and adjust feature maps.

The OFG DepthNet includes three core steps: 1) Initial depth estimation, 2) Optical flow modulation, and 3) Dual fusion interactive refinement. The initial depth features D_i are obtained by performing the SE operation on f_r and a sequence of convolutional operations, including BasicBlock, Atrous Spatial Pyramid Pooling (ASPP), and Deformable Convolutional Network (DCN). Similarly, we use the optical flow features f_o as input, passing them through the same network structure to obtain the optical flow features f_f.

Meanwhile, we perform downsampling and one-hot encoding on D_l predicted by pretrained depth estimation models to obtain sampled features D_s with consistent dimensions. Considering the inaccuracy of the depth estimation results from the depth estimation model, especially at the edges of moving objects, we use f_f to modulate D_s. The modulation process initially predicts coordinate offsets based on the spatial discrepancies between f_f and D_s using a 2D convolutional layer. Subsequently, it warps D_s using the computed flow field Ω. The main steps include: 1) Creating a mesh grid, 2) Normalizing the flow to a small range to prevent excessive warping, 3) Calculating dynamic adjustment factors, and 4) Grid clamping and bilinear sampling. The warped features are denoted as D_w.

Next, we concatenate f_f and D_w and pass them through a 3×3.2D convolutional network for feature fusion, resulting in D_k. Then, we use an adaptive integration function to output the modulated depth features D_m. The adaptive function primarily generates an attention weight map through global average pooling and 1×1.2D convolution, multiplying it with the original feature map to adaptively enhance the features. To effectively fuse the two different forms of depth features, D_i and D_m, we employ a dual fusion cross-attention module to enhance the accuracy of depth estimation. This dual fusion cross-attention module works primarily by exchanging queries during the interaction process of the two depth features.

$$\hat{D}_i = \textbf{Cross-LG-Att}.\,(D_i, D_m, D_m, d_n)\,. \tag{1}$$

$$\hat{D}_m = \textbf{Cross-LG-Att}.\,(D_m, D_i, D_i, d_n)\,. \tag{2}$$

Note that during the interaction process, we do not use the widely adopted cross-attention [65]. Considering efficiency and effectiveness, we employ Cross-LG-Attention, which is composed of neighborhood attention [66] and detailed neighborhood attention [67], with the dilation d_n set to 1 and 2, respectively, incorporating both local and global features. Next, we use Self-LG-Attention to refine and enhance the fused features, resulting in the final depth estimation distributions D_d.

D. Depth Ambiguity-Mitigated Feature Lifting

In the SSC task, effective 2D-to-3D feature lifting is crucial. Existing methods based on Lift-Splat or 2D attention mechanisms either use estimated depth to obtain pseudo-LiDAR features and map them to 3D space, which is a single-step operation without feature refinement, or they disregard depth and enhance features using 2D deformable cross-attention mechanisms. While this can provide finer semantic descriptions, it suffers from depth ambiguity issues. In contrast to these approaches, we propose a Depth Ambiguity-Mitigated Feature Lifting (DAMFL) strategy for feature transformation.

We generate extended 3D feature maps f_{3d} by performing the outer product operation between the estimated depth distributions D_s and contextual features f_c. These extended features are then converted to voxel features V_{vs} through coordinate transformation T_{camego}, post-processing and voxel pooling operations. Here, the T_{camego} represents the transformation from the camera coordinate system to the ego vehicle coordinate system. To facilitate subsequent feature querying and updating, we initialize the voxel query V_Q and combine it with V_{vs} to generate new query features V_{vq}. To further promote computational efficiency, inspired by VoxFormer [2], we utilize the pretrained depth estimation model to generate prior mask indices.

Specifically, we back-project the predicted depth maps D_l into point clouds and voxelize them to obtain binary query proposals V_{vp}. By inspecting these proposals V_{vp}, we determine the unmasked and masked indices. Using these indices, we obtain seed query features V_s from V_{vq}. Then, we adopt 3D deformable cross-attention [21] to achieve depth ambiguity-free feature lifting, which effectively addresses the inherent depth ambiguity problem in projection methods [2], [17].

More precisely, we use V_s as the query, with the extended 3D features f_{3d} as the key and value. For the query q_s in V_s located at (x, y, z) in the ego vehicle coordinate system, its reference point $R_s = (u, v, d)$ in the 3D pixel coordinate system can be calculated using formulas (3) and (4).

$$d\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{3}$$

$$u = f_x\frac{x}{d} + u_0, v = f_y\frac{y}{d} + v_0, d = z, \tag{4}$$

where f_x,f_y,u_0,v_0 are the camera intrinsic parameters. We then perform 3D deformable cross-attention in the pixel coordinate system and obtain the updated $\hat{q}_s$ as follows.

$$\hat{q}_s = \mathrm{DCA}_{3D}\,(q_s, f_{3d}) = \sum_{k=1}^{K} A_k W T_{rili}(f_{3d}, R_s + \Delta R_k). \tag{5}$$

Herein, K denotes the number of sampling points. W represents learnable weights. A denotes the attention weights. The ΔR indicates the sampling offsets. $T_{rili}(\cdot)$ represents the trilinear interpolation in f_{3d}. To further achieve refinement of the entire scene, we follow [2], [17] and employ the deformable self-attention to diffuse voxel features F_{3D}, obtaining the updated features denoted as $\hat{F}_{3D}$.

$$\hat{F}^{3D} = \mathrm{DSA}\left(F^{3D}\right) = \sum_{k=1}^{K} A_k W T_{rili}\left(\mathbf{F}^{3D}, R + \Delta R_k\right). \tag{6}$$

In doing so, we effectively lift the 2D features into the 3D space, thereby enhancing the model's geometric perception capability in scene reconstruction.

E. Geometric and Semantic Enhanced Prediction

Since the core of SSC is to predict complete geometric details and accurate semantic segmentation results, we customize two submodules to achieve Geometric and Semantic Enhanced Prediction (GSEP), rather than using simple convolutions [2], [68] or direct interpolation [12] for prediction.

To enhance geometric details, we design a lightweight residual voxel network with residual voxel blocks as the basic units. Each residual voxel block consists of two 3D convolution layers with a stride of 2, each followed by a

Batch Normalization (BN) layer and a ReLU activation function. To improve computational efficiency and save memory, we optionally employ the checkpointing strategy during the processing of each layer. In our experiments, we utilize three consecutive residual voxel blocks, denoted by m=3.

Inspired by the proven effectiveness of UNet [23] in semantic segmentation, we tailor a sparse UNet for accurate semantic perception in scenes characterized by objects and regions of varying scales. This includes a downsampling module, an Adaptive Hierarchical Aggregator (AHA) [24], an upsampling module and an interpolation part. The downsampling module utilizes the sparse convolution for feature downsampling. The AHA adopts the multi-scale Adaptive Relational convolution (AR conv), adaptive aggregator and submanifold convolution to accommodate the varying scales of different objects and backgrounds in 3D scenes. The key insight is that smaller objects like persons and cars require smaller receptive fields to capture fine semantic information, while larger areas like roads and buildings need larger receptive fields for consistent semantic reasoning. The upsampling module comprises linear and normalization operations and the interpolation part uses trilinear interpolation to adjust the results to the target size.

F. Loss Function

In this work, we follow previous work [2], [8] and treat the SSC as a voxel-wise classification problem. Adhering to the method [8], we adopt the scene-class affinity loss L_{scal} for geometric and semantic optimization. Moreover, following VoxFormer [2], we employ a weighted cross-entropy loss L_{wce} to train the network, which adjusts class weights to mitigate the impact of class imbalance. Besides, following [4], [25], our OFG DepthNet also incorporates explicit depth supervision L_{depth} during training. Overall, our loss function is defined as follows:

$$L = L_{scal}^{geo} + L_{scal}^{sem} + L_{wce} + \lambda L_{depth}. \tag{7}$$

Here, L_{scal}^{geo} and L_{scal}^{sem} are geometric and geometric scene-class affinity losses, respectively. λ represents the hyperparameter.

IV. Experiments

A. Experimental Setup

1) Dataset and Metric: SemanticKITTI [50] is the first benchmark dataset for outdoor semantic scene completion, containing LiDAR scans and front camera images from 22 sequences of the odometry benchmark of the KITTI Vision Benchmark [74]. In the evaluation setup of the SSC task, the ground truth is obtained by merging multiple LiDAR point cloud frames based on the semantic labels of the point clouds and constructing them into $256 \times 256 \times 32$ voxel grids with a resolution of $0.2m$. These grids are annotated with 19 semantic categories and 1 empty category. Following the settings in [2], sequences 0-7 and 9-10 serve as the training set, sequence 8 as the validation set, and sequences 11-22 as the test set. The images captured by the left camera are used for RGB feature extraction.

SSCBench-KITTI-360 [75] provides 7 sequences for training, 1 sequence for validation, and 1 sequence for testing, derived from the KITTI-360 [76] odometry benchmark. The training set consists of 8,487 frames from scenes 00, 02-05, 07, and 10, while the validation set includes 1,812 frames from scene 06. The testing set contains 2,566 frames from scene 09. The dataset features 19 unique semantic classes (18 semantic classes and 1 free class), with input RGB images at a resolution of 1408×376.

Occ3D-nuScenes [77], which is based on the large-scale nuScenes benchmark, provides occupancy labels for 18 classes, including 1 free class and 17 semantic classes. Out of the 1000 annotated driving scenes, 700 are allocated to the training set, 150 to the validation set, and 150 to the test set. The dataset covers a spatial range of $-40m$ to $40m$ along the X and Y axes, and $-1m$ to $5.4m$ along the Z axis. The semantic occupancy labels are defined using voxels with dimensions of $0.4m \times 0.4m \times 0.4m$ for the 17 semantic categories.

Following [8], we adopt Intersection over Union (IoU) and mean Intersection over Union (mIoU) to evaluate the effectiveness of semantic scene completion. Ideally, higher values indicate better performance. The IoU and mIoU are as follows.

$$IoU = \frac{N_{TP}}{N_{TN} + N_{FP} + N_{FN}}, \tag{8}$$

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{N_{TP_c}}{N_{TN_c} + N_{FP_c} + N_{FN_c}}, \tag{9}$$

where N_{TP} represents the "True Positive", N_{TN} indicates the "True Negative", N_{FP} stands for the "False Positive", N_{FN} denotes the "False Negative", C denotes the total number of classes.

2) Implementation Details: We train our model with PyTorch on the SemanticKITTI dataset for 25 epochs. Depth maps and optical flow images are obtained using pre-trained MobileStereoNet [28] and GMFlow [63] models, respectively. The image backbone is EfficientNet-B7 [64]. The 3D feature volume for view transformation is consistent with OccFormer [9], sized at $128 \times 128 \times 16$, with the final prediction being $256 \times 256 \times 32$, matching the ground truth. We adopt the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and set the initial learning rate to 0.0001, accompanied by a decay of 0.01. All models are trained with a batch size of 4 on 4 Nvidia 3090 GPUs, each utilizing approximately 23GB. During training, we employ the checkpointing strategy to save memory.

B. Performance on SemanticKITTI

1) Quantitative Results: As shown in Tab. II, we report the quantitative results of our method compared to the latest camera-based SSC methods on the SemanticKITTI validation set. Our approach not only significantly enhances performance in both geometric completion and semantic segmentation, as evidenced by improvements in IoU and mIoU, but also excels in critical categories crucial for autonomous driving safety. Specifically, our method surpasses Symphonies [17] across multiple categories, achieving IoU improvements of 10.33 for road, 5.74 for car, and 4.12 for motorcycle. Compared to the concurrent work CGFormer, our method achieves improvements of 0.32 and 0.33 in overall IoU and mIoU, respectively.

TABLE II

QUANTITATIVE RESULTS ON THE SEMANTICKITTI VALIDATION SET, IN TERMS OF IOU AND MIOU METRICS. † REPRESENTS THE RESULT OBTAINED USING ONLY A SINGLE IMAGE. * REPRESENTS THE REPRODUCED RESULTS IN [19]. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN BOLD

Method	SC	SSC																			
	IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-ground (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-vehicle (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traffic-sign (0.08%)	mIoU
3DSketch [CVPR20] [49]	33.30	41.32	21.63	0.00	0.00	14.81	18.59	0.00	0.00	0.00	0.00	19.09	0.00	26.40	0.00	0.00	0.00	0.73	0.00	0.00	7.50
AICNet [CVPR20] [69]	29.59	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00	8.32
MonoScene [CVPR22] [8]	37.12	57.47	27.05	15.72	0.87	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48	11.50
TPVFormer [CVPR23] [7]	35.61	56.50	25.87	20.60	0.65	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52	11.35
VoxFormer-S† [CVPR23] [2]	44.02	54.76	26.35	15.50	0.70	17.65	25.79	5.63	0.59	0.51	3.77	24.39	5.08	29.96	1.78	3.32	0.00	7.64	7.11	4.18	12.35
NDS-Scene [ICCV23] [70]	37.24	59.20	28.24	21.42	**1.67**	14.94	26.26	14.75	1.67	2.37	7.73	19.09	3.51	31.04	3.60	2.74	0.00	6.65	4.53	2.73	12.70
OccFormer [ICCV23] [9]	36.50	58.85	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	32.62	2.78	2.82	0.00	5.61	4.26	2.85	13.46
SparseOcc [CVPR24] [15]	36.48	59.59	29.68	20.44	0.47	15.41	24.03	18.07	0.78	0.89	8.94	18.89	3.46	31.06	**3.68**	0.62	0.00	6.73	3.89	2.60	13.12
Scribble2Scene [IJCAI24] [71]	43.80	49.90	29.93	20.12	0.87	20.14	24.16	17.32	0.30	1.01	3.69	25.99	8.03	32.39	1.98	0.47	0.00	6.12	7.52	5.25	13.27
HASSC-S† [CVPR24] [16]	44.82	57.05	28.25	15.90	1.04	19.05	27.23	9.91	0.92	0.86	5.61	25.48	6.15	32.94	2.80	**4.71**	0.00	6.58	7.68	4.05	13.48
H2GFormer-S† [AAAI24] [3]	44.57	56.08	29.12	17.83	0.45	19.74	27.60	10.00	0.50	0.47	7.39	26.25	6.80	34.42	1.54	2.88	0.00	7.24	7.88	4.68	13.73
OccGen [ECCV24] [18]	36.87	61.28	28.30	20.42	0.43	14.49	26.83	15.49	1.60	2.53	12.83	20.04	3.94	32.44	3.20	3.37	0.00	6.94	4.11	2.77	13.74
OctOcc [AAAI24] [12]	44.02	55.10	27.90	22.60	0.50	20.30	27.80	6.00	2.60	2.00	6.80	26.60	6.80	33.80	2.70	0.00	0.00	8.90	9.30	5.60	14.59
Symphonies [CVPR24] [17]	41.92	56.37	27.58	15.28	0.95	21.64	28.68	20.44	2.54	2.82	**13.89**	25.72	6.60	30.87	3.52	2.24	0.00	8.40	9.57	5.76	14.89
CGFormer* [NeurIPS24] [19]	45.53	65.67	32.64	24.94	0.14	21.65	32.87	9.13	3.05	2.43	9.07	**27.07**	9.07	39.35	2.44	2.71	0.00	**9.93**	12.37	6.74	16.40
Ours	**45.85**	**66.70**	**33.22**	**25.59**	0.03	**22.49**	**34.42**	4.40	**4.96**	**6.94**	11.36	26.76	**9.12**	**40.24**	2.89	2.15	0.00	9.59	**12.69**	**7.88**	**16.77**

TABLE III

QUANTITATIVE RESULTS ON THE SEMANTICKITTI TEST SET, IN TERMS OF IOU AND MIOU METRICS. † REPRESENTS THE RESULT OBTAINED USING ONLY A SINGLE IMAGE. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN BOLD

Method	SC	SSC																			
	IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-ground (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-vehicle (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traffic-sign (0.08%)	mIoU
3DSketch [CVPR20] [49]	26.85	37.70	19.80	0.00	0.00	12.10	17.10	0.00	0.00	0.00	0.00	12.10	0.00	16.10	0.00	0.00	0.00	3.40	0.00	0.00	6.23
AICNet [CVPR20] [69]	23.93	39.30	18.30	19.80	1.60	9.60	15.30	0.70	0.00	0.00	0.00	9.60	1.90	13.50	0.00	0.00	0.00	5.00	0.10	0.00	7.09
MonoScene [CVPR22] [8]	34.16	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	2.80	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10	11.08
TPVFormer [CVPR23] [7]	34.25	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	3.50	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50	11.26
VoxFormer-S† [CVPR23] [2]	42.95	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	12.20
NDS-Scene [ICCV23] [70]	36.19	58.12	28.05	25.31	6.53	14.90	19.13	4.77	1.93	2.07	**6.69**	17.94	3.49	25.01	3.44	2.77	1.64	12.85	4.43	2.96	12.58
OccFormer [ICCV23] [9]	34.53	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	12.32
Scribble2Scene [IJCAI24] [71]	42.60	50.30	20.60	27.30	11.30	23.70	20.10	**5.60**	2.70	1.60	4.50	7.90	9.60	23.80	1.40	0.00	0.00	11.30	6.50	6.60	13.33
H2GFormer-S† [AAAI24] [3]	44.20	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	13.72
Symphonies [CVPR24] [17]	42.19	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	15.04
LOMA [AAAI25] [20]	43.01	57.98	31.80	32.16	9.47	25.28	24.88	4.08	1.74	1.68	6.36	**25.63**	8.71	24.72	1.41	1.74	0.64	16.84	6.53	6.05	15.10
Ours	**45.56**	**62.70**	**34.10**	**34.20**	**12.00**	**26.30**	**25.90**	4.70	**3.80**	**3.10**	2.40	**24.00**	**10.20**	**27.70**	1.50	**3.10**	0.80	**17.40**	**8.20**	**9.10**	**16.39**

Notably, our method demonstrates significant gains in the car and motorcycle categories. Moreover, it consistently excels on the SemanticKITTI test set, as depicted in Fig. 1. We also provide the complete category comparison, as shown in Tab. III.

2) Qualitative Results: Fig. 3 presents a qualitative comparison of our method with previous state-of-the-art methods, including Monoscene [8], VoxFormer [2], OccFormer [9] and Symphonies [17]. We observe that our method has the following advantages over other methods: 1) It avoids the semantic confusion observed in VoxFormer and Symphonies, such as misclassifications between buildings and terrain; 2) It produces clearer boundaries in the car category, avoiding trailing artifacts, as demonstrated in the first row of our results; 3) It generates spatial layouts that more accurately match the ground truth, evident in the third row for roads; 4) It successfully detects persons in front of cars, a task where other methods fail, as shown in the second row. These results further validate the effectiveness of our approach.

C. Performance on SSCBench-KITTI-360 and Occ3D-nuScenes

We also perform a quantitative analysis on the SSCBench-KITTI360 dataset, with the experimental results presented in Tab. IV. We compare our method against eight state-of-the-art camera-based SSC methods and demonstrate superior overall performance. Specifically, compared to Symphonies, our method achieves improvements of 1.16 in mIoU and 4.49 in IoU. Notably, for categories such as car and motorcycle, our method achieves mIoU gains of 1.38 and 2.03, respectively,

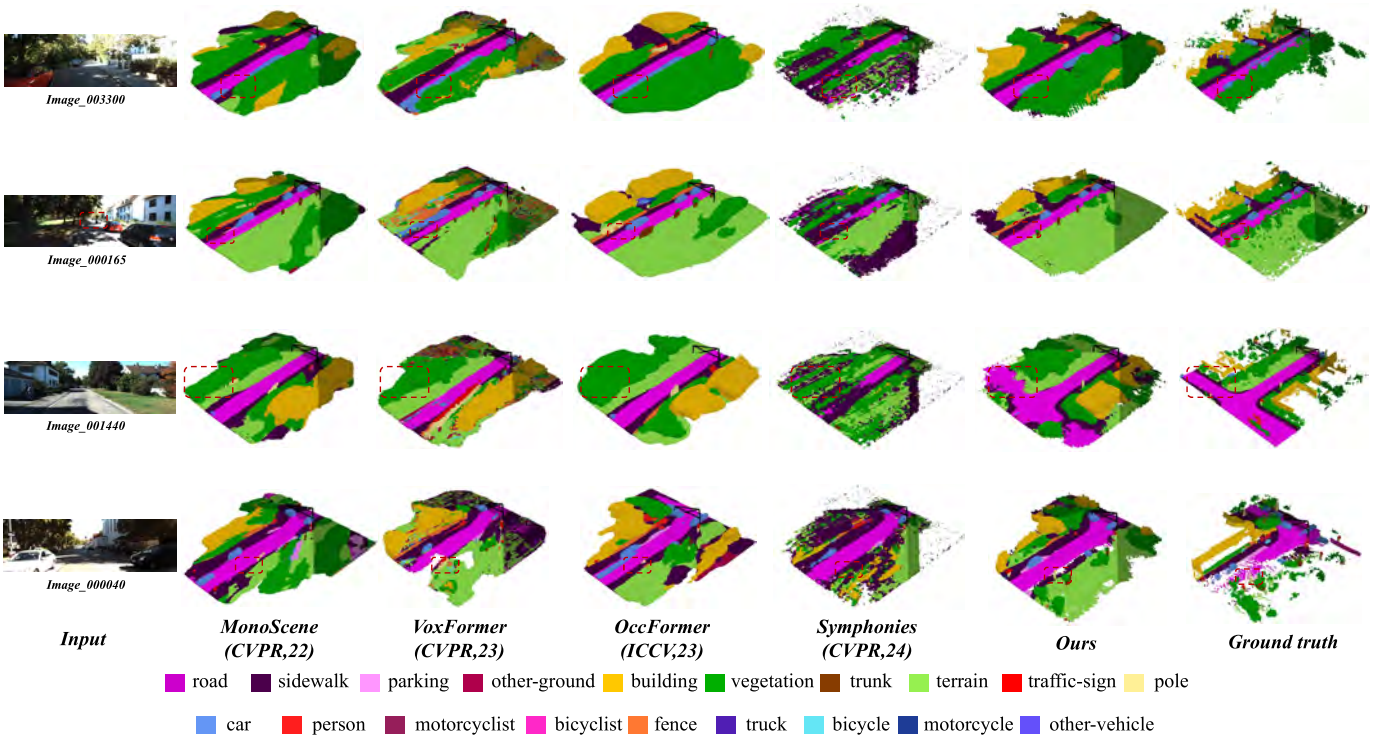

Fig. 3. Qualitative visualizations on the SemanticKITTI validation set. Our method avoids generating long traces, such as cars, and accurately detects persons in front of cars, as demonstrated in the first and second rows. Furthermore, our approach maintains a spatial layout that aligns more closely with the ground truth, as exemplified by the road in the third row.

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE OUTDOOR SCENES FROM THE SSCBENCH-KITTI-360 TEST SET IN TERMS OF IOU AND MIOU METRICS. † REPRESENTS THE RESULT OBTAINED USING ONLY A SINGLE IMAGE. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN BOLD

	SC	SSC																		
Method	IoU	road (14.98%)	sidewalk (6.43%)	parking (2.31%)	other-ground (2.05%)	building (15.67%)	car (2.85%)	truck (0.16%)	bicycle (0.01%)	motorcycle (0.01%)	other-vehicle (5.75%)	vegetation (41.99%)	terrain (7.10%)	person (0.02%)	fence (0.96%)	pole (0.22%)	traffic-sign (0.06%)	other-struct (4.33%)	other-object (0.28%)	mIoU
Monoscene [CVPR22] [8]	37.87	48.35	28.13	11.38	3.32	32.89	19.34	8.02	0.43	0.58	2.03	26.15	16.75	0.86	3.53	6.92	5.67	4.20	3.09	12.31
VoxFormer-S† [CVPR23] [2]	38.76	47.01	27.21	9.67	2.89	31.18	17.84	4.56	1.16	0.89	2.06	28.99	14.69	1.63	4.97	6.51	6.92	3.79	2.43	11.91
TPVFormer [CVPR23] [7]	40.22	52.99	31.07	11.99	3.78	34.83	21.56	8.06	1.09	1.37	2.57	30.08	17.52	2.38	4.80	7.46	5.86	5.48	2.70	13.64
OccFormer [ICCV23] [9]	40.27	54.30	31.53	13.44	3.55	36.42	22.58	9.89	0.66	0.26	3.82	31.00	19.51	2.77	4.80	7.77	8.51	6.95	4.60	13.81
IAMSSC [TITS24] [72]	41.80	50.56	15.08	29.32	3.22	40.22	25.20	10.83	0.83	0.66	2.45	33.05	19.51	7.92	5.60	4.71	5.97	3.09	1.45	12.97
DepthSSC [WACV25] [73]	40.85	50.80	15.85	32.32	4.85	40.08	25.02	10.56	0.66	0.75	2.66	33.08	21.13	7.97	5.75	5.05	5.97	3.71	2.50	14.28
Symphonies [CVPR24] [17]	44.12	54.94	13.83	32.76	6.93	35.11	30.02	11.24	1.25	1.56	5.90	**38.33**	**25.11**	14.01	8.58	14.44	9.57	11.28	2.20	18.58
LOMA [AAAI25] [20]	46.35	58.00	37.52	15.76	**7.51**	**41.20**	27.59	11.49	**2.57**	3.57	7.47	37.72	20.27	5.53	8.42	14.62	**16.40**	10.49	6.51	18.28
Ours	**48.61**	**63.40**	**37.62**	**34.34**	5.08	34.43	**30.40**	**12.59**	1.97	**3.59**	**9.48**	32.10	20.35	**14.29**	**9.79**	**17.00**	9.73	**11.83**	**7.32**	**19.74**

further validating its enhanced perception capabilities for dynamic objects. In comparison to the latest method, LOMA, our approach also achieves improvements of 1.46 in mIoU and 2.26 in IoU. Additionally, significant performance gains are observed in the road and sidewalk categories. We attribute this improvement to the effectiveness of our method in mitigating ambiguity issues arising in the 3D-2D depth projection process.

Moreover, our method demonstrates excellent performance on the semantic occupancy prediction dataset Occ3D-nuScenes, as shown in Tab. V. We compare our method with five other classic approaches. Compared to the classic method TPVFormer, our method achieves an improvement of 5.22 mIoU. Notably, our method not only achieves the best results in key categories (such as car and pedestrian) but also performs excellently in background categories (such as terrain and vegetation), with a particularly significant gain of 13.4 in the vegetation category.

D. Ablation Studies

All experiments are conducted on the SemanticKITTI validation set unless otherwise stated.

1) Analysis of Optical Flow-Guided DepthNet: We conduct ablation studies on the OFG DepthNet component. Setting A, serving as the baseline, adopts the DepthNet from DFA3D [21], removes prior mask indices in feature lifting, and

TABLE V

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE OCC3D-NUSCENES DATASET. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN BOLD

Method	Supervision	mIoU	others	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [CVPR22] [8]	3D	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
OccFormer [ICCV23] [9]	3D	21.93	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97
BEVFormer [TPAMI24] [78]	3D	26.88	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.0	28.06	20.04	17.69
CTF-Occ [NeurIPS23] [77]	3D	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.0
TPVFormer [CVPR23] [7]	3D	27.83	7.22	38.90	13.67	40.78	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	35.47	37.55	30.70	19.40	16.78
Ours	3D	**33.05**	**8.82**	**40.10**	**22.26**	**42.34**	**48.12**	**18.90**	**24.98**	**23.85**	**22.02**	**28.76**	**35.36**	**61.65**	**40.28**	**42.72**	**42.47**	**28.88**	**30.18**

TABLE VI

ABLATION STUDY OF THE ARCHITECTURAL COMPONENTS. "RESIDUAL VOXEL NETWORK" IS ABBREVIATED AS "RES." AND "SPARSE UNET" IS SHORTENED TO "SPA."

Setting	OFG DepthNet DFIR	OFG DepthNet OFM	DAMFL Prior m.	GSEP Res.	GSEP Spa.	Training Memory	Inference Time	IoU	mIoU
A						**17.8 G**	**199 ms**	44.08	14.98
B	✓					20.0 G (+2.2 G)	208 ms (+9 ms)	44.57(+0.49)	15.27(+0.29)
C	✓	✓				20.4 G (+0.4 G)	212 ms (+4 ms)	44.73(+0.16)	15.34(+0.07)
D	✓	✓	✓			20.7 G (+0.3 G)	214 ms (+2 ms)	45.68(+0.95)	16.03(+0.69)
E	✓	✓	✓	✓		21.5 G (+0.8 G)	217 ms (+3 ms)	45.82(+0.14)	16.41(+0.38)
F	✓	✓	✓	✓	✓	23.2 G (+1.7 G)	227 ms (+10 ms)	**45.85**(+0.03)	**16.77**(+0.36)

TABLE VII

COMPARATIVE ANALYSIS OF DFIR SETTINGS AND THE EFFECTS OF PARAMETER m

Ablation on DFIR settings			Ablation on the parameter m		
Setting	IoU	mIoU	m	IoU	mIoU
w/o neigh. atten.	44.26	15.09	1	45.69	16.20
w/ neigh. atten.	44.39	15.23	2	45.70	16.32
w/ LG-atten.(d_n=1,2)	**44.57**	**15.27**	3	**45.82**	16.41
w/ LG-atten.(d_n=1,3)	44.54	15.21	4	45.81	**16.42**

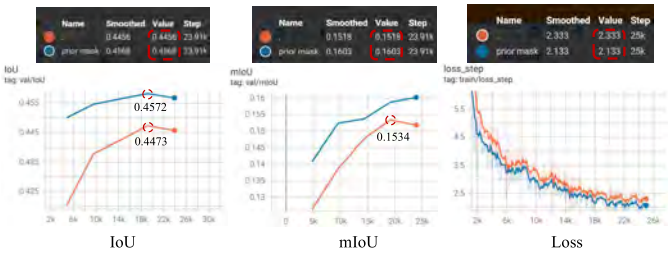

Fig. 4. Analysis of the effectiveness of prior mask indices. The values of IoU and mIoU are scaled down by a factor of 100.

eliminates enhanced geometric and semantic subnetworks. Setting B introduces Dual Fusion Interactive Refinement (DFIR) module. Setting C further adds the Optical Flow Modulation (OFM) module. As shown in Tab. VI, the OFG DepthNet improves IoU and mIoU by 0.65 and 0.36, respectively, compared to the baseline, demonstrating its effectiveness. Tab. VII presents a detailed ablation study of DFIR settings, demonstrating the effectiveness of using LG-Attention.

2) Analysis of Depth Ambiguity-Mitigated Feature Lifting: In the baseline, we use 3D deformable attention [21] and deformable self-attention without any other modifications, which surpasses existing methods such as VoxFormer [2]. This demonstrates that 3D deformable attention performs better than 2D deformable attention in the SSC task by avoiding depth ambiguities during the projection process.

Apart from the aforementioned advantage, our method further benefits from the integration of prior mask indices into the Depth Ambiguity-Mitigated Feature Lifting (DAMFL) component, which results in a 0.95 increase in IoU and a 0.69 increase in mIoU, as shown in Tab. VI. Fig. 4 further illustrates the variations in IoU and mIoU (scaled by 100) on the SemanticKITTI validation set, as well as the loss values. The results clearly indicate that the incorporation of prior mask indices significantly enhances both IoU and mIoU, while also achieving the reduction in loss values, thereby confirming the effectiveness of prior mask indices in improving model performance.

3) Analysis of Geometric and Semantic Enhanced Prediction: We conduct ablation studies on the Geometric and Semantic Enhanced Prediction (GSEP) component and report the results in Tab. VI. As evidenced by our experiments, introducing the residual voxel network results in better performance on all metrics in setting E. Additionally, we examine the impact of varying the number of residual voxel blocks, parameter m, as shown in Tab. VII. To strike a balance between memory usage and computational efficiency, we chose $m = 3$. In setting F, further utilizing the sparse UNet yields improvements of 0.03 in IoU and 0.36 in mIoU.

4) Generalization Analysis Across Different Pre-Trained Estimation Models: To verify the generalization ability of our method across different pre-trained models, we evaluate several depth estimation and optical flow models. For depth estimation, we compare stereo-based models (including MobileStereoNet [62], MonSter [79], and Stereo Anywhere [80]) with monocular models (such as Adabins [81] and

TABLE VIII

GENERALIZATION COMPARISON ACROSS DIFFERENT PRE-TRAINED MODELS. (S) DENOTES METHODS BASED ON STEREO IMAGE PAIR INPUT; (M) INDICATES MONOCULAR SETTINGS

Method	Different depth estimation model					Different optical flow estimation model		
	MobileStereoNet (S)	MonSter (S)	Stereo Anywhere (S)	Adabins (M)	Diffusiondepth (M)	GMFlow	FlowFormer	StreamFlow
IoU	45.85	45.87	**45.92**	43.79	45.06	45.85	45.83	**45.88**
mIoU	16.77	16.80	**16.84**	15.69	16.33	16.77	16.74	**16.81**

TABLE IX

ANALYSIS OF SINGLE-VIEW SETUP

Setting	IoU	mIoU
- (baseline)	**45.85**	**16.77**
w/o optical flow	45.71	16.72
w/o optical flow, w/ DiffusionDepth	44.84	16.25

TABLE X

ANALYSIS OF DIFFERENT BACKBONE NETWORKS

Setting	IoU	mIoU
EfficientNetB7	**45.85**	**16.77**
ResNet50	45.84	16.69

TABLE XI

COMPARISON OF MEMORY USAGE AND INFERENCE TIME

Method	Memory	Inference.	IoU	mIoU
VoxFormer-S† [2]	**15.2G**	208 ms	44.02	12.35
OccFormer [9]	17.9G	**201 ms**	36.50	13.46
StereoScene [14]	20.0G	262 ms	43.85	15.43
Ours	23.2G	227 ms	**45.85**	**16.77**

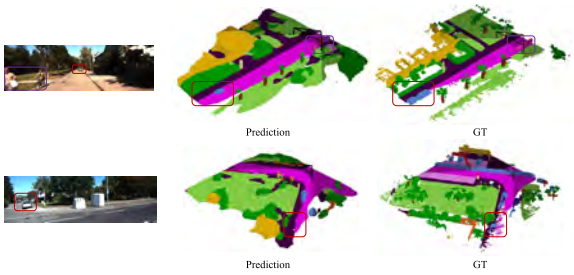

Fig. 5. Failure cases.

DiffusionDepth [82]). As shown in Tab. VIII, using Stereo Anywhere as the depth pre-trained model further improves the performance of our method, thanks to its stronger robustness and zero-shot generalization ability. Moreover, we notice that, when using monocular setups, the performance of our method decreases slightly compared to stereo methods, but it remains competitive. Specifically, when using Diffusiondepth, our method only decreases by 0.34 mIoU and 0.79 IoU compared to the baseline. We also evaluate the generalization ability of our method on three different optical flow models: GMFlow [63] (baseline), FlowFormer [83], and StreamFlow [84]. The experimental results show that using StreamFlow further improves performance.

5) Analysis of Single-View Setup: We also perform ablation experiments on the single-view setup, which include the following three configurations: 1) the original method setup as the baseline; 2) removing the motion information from optical flow images to analyze its impact on depth information gain; 3) replacing MobileStereoNet with DiffusionDepth based on configuration 2) to evaluate performance under the single-view setup. According to the results in Tab. IX, after removing the modulation of depth information by optical flow images, the mIoU decreases by 0.05, which is expected, as the gain from optical flow in depth estimation is limited. Additionally, we observe that, based on configuration 2), using monocular DiffusionDepth leads to a decrease of 0.47 in mIoU. This suggests that stereo-based depth estimation models still outperform monocular depth estimation models. However, overall, our method still achieves a competitive mIoU of 16.25 under the single-view setup.

6) Analysis of Different Backbone Networks: We compare the two most commonly used backbone networks (ResNet50 and EfficientNetB7) in SSC tasks. Tab. X shows that both backbones maintain good performance for our method. Compared to ResNet50, using EfficientNetB7 results in a 0.01 improvement in IoU and a 0.08 improvement in mIoU.

E. Computational Cost

We compare our method with classical approaches in terms of computational memory and inference time. Tab. XI reveals that while our method marginally increases memory usage, the inference time remains moderate, offering opportunities for future optimizations. Furthermore, as shown in Tab. VI, we also provide a step-by-step analysis of the impact of each core module on training memory and inference time. Specifically, the DFIR module introduces approximately 2.2 GB of memory overhead and 9 ms of inference latency, while the Sparse UNet module adds 1.7 GB of memory usage and 10 ms latency. This fine-grained evaluation provides clear guidance for further optimization of the method.

F. Failure Cases

As shown in Fig. 5, our method still exhibits certain failure cases. In the purple solid box of the first row, the "bicyclist" category is misclassified as "person." This issue can be primarily attributed to the highly imbalanced category distribution in our training data: the "bicyclist" and "person" categories each represent only 0.07% of the dataset, while the "road" and "vegetation" categories account for 15.30% and 39.3%, respectively. This significant class imbalance results in insufficient learning of minority categories, which is particularly

critical for safety-sensitive perception tasks in autonomous driving. Additionally, our model fails to detect cars when their regions are heavily occluded in the input image, as shown in the red dashed box. These results highlight the limitation of using front-view inputs alone in complex scenarios, as they may not provide comprehensive situational awareness. This further emphasizes the need to integrate surrounding-view sensor data or millimeter-wave radar, which offers a cost-effective alternative to LiDAR, especially in occlusion-prone environments.

G. Discussion

We believe there are still several points worth exploring and further investigation.

1) Our method aims to improve the accuracy of depth estimation by incorporating additional optical flow and depth images. Based on both empirical evidence and experimental results, this approach proves to be effective, consistent with prior studies that enhance performance by introducing additional depth [2] and segmentation maps [72]. However, few studies have thoroughly investigated whether such performance improvements primarily result from the extra data extracted by pre-trained models, leading to gains attributed to increased data volume. This issue warrants further exploration.
2) In the current SemanticKITTI dataset, the inadequate handling of motion blur artifacts introduces labeling errors in the ground truth. For example, as shown in the first row of Fig. 2 (car), these labeling inaccuracies directly affect the reliability of quantitative evaluations. While our method effectively avoids erroneous motion trajectories in visualized results and provides reasonable predictions, the presence of inaccurate labels significantly reduces the measured improvements in the evaluation tables. This further highlights the importance and necessity of self-supervised methods to mitigate the dependency on manual labeling.
3) Although existing methods, such as those based on NeRF [60] and 3DGS [61], reduce the high cost of 3D annotations through self-supervised strategies, they still face notable limitations. NeRF suffers from slow training speeds, while 3DGS struggles to represent scene geometry with sufficient accuracy. Additionally, a commonly overlooked issue is that camera-based semantic scene completion methods predominantly rely on monocular or stereo RGB image inputs. With a limited number of viewpoints, these methods must effectively address the challenges associated with sparse-view scenarios.

V. Conclusion

This paper introduces a comprehensive enhanced framework for Semantic Scene Completion (SSC). To address challenges such as depth errors and depth ambiguities during the 2D to 3D transformation process, we integrate the OFG DepthNet with a depth ambiguity-mitigated feature lifting strategy to enhance prediction accuracy in regions with notable depth changes and simultaneously avoid depth ambiguities. Additionally, we customize two subnetworks: a residual voxel network and a sparse UNet, to further enhance the network's geometric prediction capabilities and secure consistent semantic reasoning across varying scales. We conduct extensive experiments and ablation studies to demonstrate the superiority of the proposed method on the SemanticKITTI, SSCBench-KITTI-360 and Occ3D-nuScene datasets.

References

[1] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–18.

[2] Y. Li et al., "VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9087–9098.

[3] Y. Wang and C. Tong, "H2GFormer: Horizontal-to-global voxel transformer for 3D semantic scene completion," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 5722–5730.

[4] X. Wang et al., "OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 17850–17859.

[5] B. Li et al., "Hierarchical temporal context learning for camera-based semantic scene completion," 2024, *arXiv:2407.02077*.

[6] L. Zhao et al., "LowRankOcc: Tensor decomposition and low-rank recovery for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 9806–9815.

[7] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9223–9232.

[8] A.-Q. Cao and R. de Charette, "MonoScene: Monocular 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3991–4001.

[9] Y. Zhang, Z. Zhu, and D. Du, "OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9433–9443.

[10] Z. Li et al., "FB-OCC: 3D occupancy prediction based on forward-backward view transformation," 2023, *arXiv:2307.01492*.

[11] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 194–210.

[12] W. Ouyang, X. Song, B. Feng, and Z. Xu, "OctOcc: high-resolution 3D occupancy prediction with octree," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 5, pp. 4369–4377.

[13] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4794–4803.

[14] B. Li et al., "Bridging stereo geometry and BEV representation with reliable mutual interaction for semantic scene completion," 2023, *arXiv:2303.13959*.

[15] P. Tang et al., "SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 15035–15044.

[16] S. Wang et al., "Not all voxels are equal: hardness-aware semantic scene completion with self-distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 14792–14801.

[17] H. Jiang et al., "Symphonize 3D semantic scene completion with contextual instance queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 20258–20267.

[18] G. Wang et al., "OccGen: Generative multi-modal 3D occupancy prediction for autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 95–112.

[19] Z. Yu et al., "Context and geometry aware voxel transformer for semantic scene completion," 2024, *arXiv:2405.13675*.

[20] Y. Cui, Z. Li, J. Wang, and Z. Fang, "LOMA: Language-assisted semantic occupancy network via triplane mamba," 2024, *arXiv:2412.08388*.

[21] H. Li et al., "DFA3D: 3D deformable attention for 2D-to-3D feature lifting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6661–6670.

[22] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.

[23] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 205–218.

[24] B. Peng et al., "OA-CNNs: Omni-adaptive sparse CNNs for 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 21305–21315.

[25] Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1477–1485.

[26] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13760–13769.

[27] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.

[28] F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, "MobileStereoNet: Towards lightweight deep networks for stereo matching," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 677–686.

[29] R. Charles, H. Su, K. Mo, and L. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 77–85.

[30] X. Ren, J. Huang, X. Zeng, K. Museth, S. Fidler, and F. H. Williams, "XCube: Large-scale 3D generative modeling using sparse voxel hierarchies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 4209–4219.

[31] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12919–12928.

[32] S. Cheng, X. Chen, X. He, Z. Liu, and X. Bai, "PRA-Net: Point relation-aware network for 3D point cloud analysis," *IEEE Trans. Image Process.*, vol. 30, pp. 4436–4448, 2021.

[33] D. Li, K. Ma, J. Wang, and G. Li, "Hierarchical prior-based super resolution for point cloud geometry compression," *IEEE Trans. Image Process.*, vol. 33, pp. 1965–1976, 2024.

[34] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[35] S. Zhou et al., "Feature 3DGS: Supercharging 3D Gaussian splatting to enable distilled feature fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 21676–21685.

[36] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner, "Scan2CAD: Learning CAD model alignment in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2614–2623.

[37] A. Avetisyan, A. Dai, and M. Niessner, "End-to-end CAD model retrieval and 9DoF alignment in 3D scans," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2551–2560.

[38] A. Dai, C. Diller, and M. Nießner, "SG-NN: Sparse generative neural networks for self-supervised scene completion of RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 849–858.

[39] A. Dai, Y. Siddiqui, J. Thies, J. Valentin, and M. Niesner, "SPSG: Self-supervised photometric scene generation from RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1747–1756.

[40] H. Chen, J. Huang, T. Mu, and S. Hu, "CIRCLE: Convolutional implicit reconstruction and completion for large-scale indoor scene," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 506–522.

[41] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon, "High-quality depth map upsampling and completion for RGB-D cameras," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5559–5572, Dec. 2014.

[42] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3504–3515.

[43] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[44] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9650–9660.

[45] A. Mirzaei et al., "SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 20669–20679.

[46] E. Weber et al., "NeRFiller: Completing scenes via generative 3D inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 20731–20741.

[47] Z. Liu et al., "InFusion: Inpainting 3D Gaussians via learning depth completion from diffusion prior," 2024, *arXiv:2404.11613*.

[48] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 190–198.

[49] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3D sketch-aware semantic scene completion via semi-supervised structure prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4192–4201.

[50] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9297–9307.

[51] L. Rold ao, R. de Charette, and A. Verroust-Blondet, "LMSCNet: Lightweight multiscale 3D semantic completion," in *Proc. Int. Conf. 3D Vis. (3DV)*, Nov. 2020, pp. 111–119.

[52] J. Tang, X. Chen, J. Wang, and G. Zeng, "Not all voxels are equal: Semantic scene completion from the point-voxel perspective," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 2, pp. 2352–2360.

[53] Z. Xia et al., "SCPNet: Semantic scene completion on point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17642–17651.

[54] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrila, "Semantic scene completion using local deep implicit functions on LiDAR data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7205–7218, Oct. 2022.

[55] A.-Q. Cao, A. Dai, and R. De Charette, "PaSCo: Urban 3D panoptic scene completion with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14554–14564.

[56] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2023, pp. 21672–21683.

[57] J. Mei et al., "Camera-based 3D semantic scene completion with sparse guidance network," *IEEE Trans. Image Process.*, vol. 33, pp. 5468–5481, 2024.

[58] M. Wang, H. Pi, R. Li, Y. Qin, Z. Tang, and K. Li, "VLScene: Vision-language guidance distillation for camera-based 3D semantic scene completion," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 8, pp. 7808–7816.

[59] Z. An et al., "Multimodality helps few-shot 3D point cloud semantic segmentation," 2024, *arXiv:2410.22489*.

[60] A. Hayler, F. Wimbauer, D. Muhle, C. Rupprecht, and D. Cremers, "S4C: Self-supervised semantic scene completion with neural fields," in *Proc. Int. Conf. 3D Vis. (3DV)*, Mar. 2024, pp. 409–420.

[61] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "GaussianFormer: Scene as Gaussians for vision-based 3D semantic occupancy prediction," 2024, *arXiv:2405.17429*.

[62] F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, "MobileStereoNet: Towards lightweight deep networks for stereo matching," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 2417–2426.

[63] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning optical flow via global matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 8111–8120.

[64] Y. Pu, Y. Han, Y. Wang, J. Feng, C. Deng, and G. Huang, "Fine-grained recognition with learnable semantic data augmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 3130–3144, 2024.

[65] X. Yi, X. Han, H. Zhang, L. Tang, and J. Ma, "Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 27016–27025.

[66] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 6185–6194.

[67] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," 2022, *arXiv:2209.15001*.

[68] H. Xiao, W. Kang, H. Liu, Y. Li, and Y. He, "Semantic scene completion via semantic-aware guidance and interactive refinement transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 4212–4225, May 2025.

[69] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3351–3359.

[70] J. Yao et al., "NDC-scene: Boost monocular 3D semantic scene completion in normalized device coordinates space," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9421–9431.

[71] S. Wang et al., "Label-efficient semantic scene completion with scribble annotations," 2024, *arXiv:2405.15170*.

[72] H. Xiao, H. Xu, W. Kang, and Y. Li, "Instance-aware monocular 3D semantic scene completion," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6543–6554, Jul. 2024.
[73] J. Yao, J. Zhang, X. Pan, T. Wu, and C. Xiao, "DepthSSC: Monocular 3D semantic scene completion via depth-spatial alignment and voxel adaptation," 2023, *arXiv:2311.17084*.
[74] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
[75] Y. Li et al., "SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 13333–13340.
[76] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.
[77] X. Y. Tian, T. Jiang, L. Yun, Y. Wang, Y. Wang, and H. Zhao, "Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 64318–64330.
[78] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 2020–2036, Mar. 2025.
[79] J. Cheng et al., "MonSter: Marry monodepth to stereo unleashes power," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Jun. 2025, pp. 6273–6282.
[80] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 1013–1027.
[81] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jan. 2021, pp. 4009–4018.
[82] Y. Duan, X. Guo, and Z. Zhu, "DiffusionDepth: Diffusion denoising approach for monocular depth estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 432–449.
[83] Z. Huang et al., "FlowFormer: A transformer architecture for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 668–685.
[84] S. Sun, J. Liu, T. H. Li, H. Li, G. Liu, and W. Gao, "StreamFlow: Streamlined multi-frame optical flow estimation for video sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 9205–9228.

Haihong Xiao received the Ph.D. degree from South China University of Technology. From November 2023 to December 2024, he was a Visiting Student with the College of Computing and Data Science, Nanyang Technological University, Singapore, supported by the China Scholarship Council (CSC). He is currently a Lecturer with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests lie in 3D computer vision, scene representation learning, and spatial AI.

Wenxiong Kang (Member, IEEE) received the Ph.D. degree from South China University of Technology, Guangzhou, China, in 2009. He is currently a Professor with the School of Automation Science and Engineering, South China University of Technology. His research interests include computer vision, biometrics identification, image processing, and pattern recognition. In recent years, he has published more than 100 articles in the significant domestic and international journals or conferences, including European Conference on Computer Vision (ECCV), the International Conference on Computer Vision (ICCV), and the International Journal of Computer Vision (IJCV). He was a recipient of the Distinguished Paper Award at the 2021 Association Conference on Artificial Intelligence (AAAI) and the Best Paper Award at the 2021 International Joint Conference on Biometrics (IJCB).

Yulan Guo (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT) in 2008 and 2015, respectively. He has authored more than 150 papers at highly refereed journals and conferences. His research interests include 3D vision, low-level vision, and machine learning. He is a Senior Member of ACM. He also served as an Area Chair for CVPR in 2023 and 2021, ICCV 2021, ECCV 2024, NeurIPS 2024, and ACM Multimedia 2021. He organized more than ten workshops, challenges, and tutorials in prestigious conferences, such as CVPR, ICCV, ECCV, and 3DV. He served as a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and an Associate Editor for *Computers & Graphics*, *IET Computer Vision*, and *IET Image Processing*.

Hao Liu received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC) in 2016, the M.S. degree from the National University of Defense Technology (NUDT) in 2018, and the Ph.D. degree from Sun Yat-sen University in 2023. He is currently a Principal Investigator (PI) and a Zijiang Young Scholar with the School of Geospatial Artificial Intelligence, East China Normal University. From 2023 to 2025, he was a Post-Doctoral Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests lie in spatial intelligence, 3D vision, and point cloud processing, particularly in 3D object detection and tracking.

Ying He (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, China, and the Ph.D. degree in computer science from Stony Brook University, USA. He is currently an Associate Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests fall into the general areas of visual computing and he is particularly interested in the problems which require geometric analysis and computation. He actively participates in the technical program committees of major conferences in geometric modeling. He is serving/has served on the Editorial Board for IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, *Computer Graphics Forum*, and *Computational Visual Media*. He has also served as the General/Program Co-Chair for the Shape Modeling International in 2022, the Symposium on Solid and Physical Modeling in 2022 and 2023, the Geometric Modeling and Processing in 2014 and 2021, and the Conference on Computational Visual Media in 2020.