Semantic Scene Completion via Semantic-Aware Guidance and Interactive Refinement Transformer

Haihong Xiao, Wenxiong Kang, *Member, IEEE*, Hao Liu, Yuqiong Li, and Ying He, *Member, IEEE*

***Abstract*— Predicting per-voxel occupancy status and corresponding semantic labels in 3D scenes is pivotal to 3D intelligent perception in autonomous driving. In this paper, we propose a novel semantic scene completion framework that can generate complete 3D volumetric semantics from a single image at a low cost. To the best of our knowledge, this is the first endeavor specifically aimed at mitigating the negative impacts of incorrect voxel query proposals caused by erroneous depth estimates and enhancing interactions for positive ones in camera-based semantic scene completion tasks. Specifically, we present a straightforward yet effective Semantic-aware Guided (SAG) module, which seamlessly integrates with task-related semantic priors to facilitate effective interactions between image features and voxel query proposals in a plug-and-play manner. Furthermore, we introduce a set of learnable object queries to better perceive objects within the scene. Building on this, we propose an Interactive Refinement Transformer (IRT) block, which iteratively updates voxel query proposals to enhance the perception of semantics and objects within the scene by leveraging the interaction between object queries and voxel queries through query-to-query cross-attention. Extensive experiments demonstrate that our method outperforms existing state-of-the-art approaches, achieving overall improvements of 0.30 and 2.74 in mIoU metric on the SemanticKITTI and SSCBench-KITTI-360 validation datasets, respectively, while also showing superior performance in the aspect of small object generation.**

***Index Terms*— 3D vision, semantic scene completion, interactive refinement transformer.**

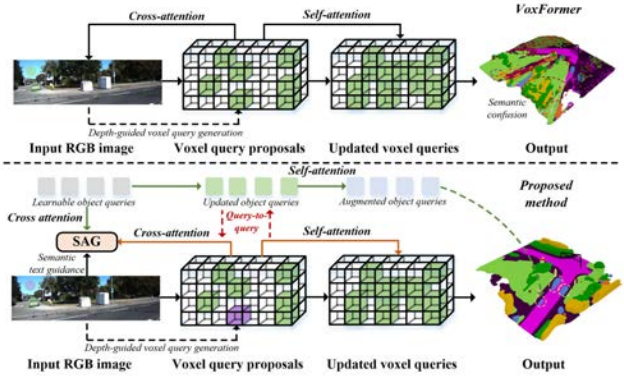

Fig. 1. **a) VoxFormer** uses a single 2D image to predict complete 3D geometry and semantics. It primarily involves voxel query proposal generation based on depth, with information interactions and enhancements facilitated by deformable cross-attention and self-attention mechanisms, respectively (mask parts are omitted for conciseness). **b) Our method** introduces a Semantic-aware Guided (SAG) module that weakens the interaction of geometrically relevant but semantically irrelevant image features associated with negative query proposals *(purple voxel)* while simultaneously enhancing the interaction between positive query proposals and areas of interest. We also introduce a set of learnable object queries and further propose the query-to-query cross attention to enhance the perception of semantics and objects within the scene.

I. Introduction

SEMANTIC Scene Completion (SSC) is a fundamental and emerging task in autonomous driving, aiming to predict per-voxel occupancy and corresponding semantic labels from partial point clouds or image inputs. Recent vision-centric SSC schemes represent a promising path toward more systematic generalization by precisely perceiving and reconstructing a fine-grained 3D world with consumer-grade RGB cameras, thereby revolutionizing the landscape of 3D intelligent perception [1], [2], [3].

Early works on SSC primarily focus on indoor scenarios, using depth maps [4], additional RGB images [5], or TSDF [6] as inputs. With the availability of outdoor scene datasets, such as SemanticKITTI [7], an increasing number of researchers from both academia and industry are now shifting their focus to more challenging outdoor scenarios. Despite substantial progress, they rely heavily on point clouds obtained from expensive LiDAR as input, which limits their widespread applicability. Furthermore, the sparsity of scanned points, especially for long-range objects, still lacks an effective feature extraction solution. Inspired by the success of BEVFormer [8] in 3D object detection, researchers have rethought the purely vision-based SSC task, leading to new efforts based on surrounding view-based [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] and monocular camera-based [29], [30], [31], [32], [33] methods.

Received 7 October 2024; revised 15 November 2024; accepted 11 December 2024. Date of publication 16 December 2024; date of current version 7 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62376100, in part by the Ministry of Education, Singapore, under its Academic Research Fund Grants (MOE-T2EP20220-0005 & RT19/22), in part by the International Science and Technology Cooperation Project of Guangzhou Economic and Technological Development District under Grant 2023GH16, and in part by the Fundamental Research Funds for the Central Universities under Grant 2024ZYGXZR104. This article was recommended by Associate Editor Z. Tang. *(Corresponding authors: Wenxiong Kang; Ying He.)*

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

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

Yuqiong Li is with the Key Laboratory for Mechanics in Fluid Solid Coupling Systems, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: liyuqiong@imech.ac.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3518493

As a pioneering work, MonoScene [29] introduces a Features Line of Sight Projection (FLoSP) module, which projects 2D features into 3D space in a depth-agnostic manner. Undoubtedly, due to the lack of effective geometric guidance, this projection inevitably maps 2D features into incorrect spatial areas, such as empty or overlapping semantic regions. To address this issue, recent efforts have sought to employ more view information, including stereo [34] and surrounding views [9], along with novel structural representations, such as the Tri-Perspective View (TPV) [18] and additional geometric priors, like depth maps [30], to enhance geometric perception capabilities. Subsequently, they utilize the cross-attention to aggregate information across different perspectives or the deformable cross-attention [35] to enable cross-modal interactions between geometric proposals and image features. In essence, the deformable cross-attention mechanism avoids considering the spatial positions of entire image pixels and focuses only on the most relevant positional information near the geometry. This not only boosts computational efficiency, but also makes the generation of complex 3D scenes feasible, serving as a stepping stone to a generic backbone for SSC.

Although surrounding views can implicitly construct geometric relationships, they significantly increase the computational burden. The TPV representation extends beyond the limitations of the single-plane Bird's Eye View (BEV) [8] representation, yet it still leads to a certain loss of geometric details. Consequently, we follow the promising avenue of incorporating additional depth priors to enhance the geometric perception capabilities of SSC. However, certain limitations impede the potential of the deformable cross-attention mechanism in achieving precise 3D semantic scene completion. **1) Negative voxel query proposal generation.** The accuracy of voxel query proposals heavily depends on the accuracy of depth estimators. Due to scale ambiguity and occlusions, current depth estimators generally produce inaccurate depth estimations, especially for some uninterested regions, resulting in negative voxel query proposals that greatly impact final SSC performance. **2) Category-agnostic feature extraction.** Existing image feature extractors for the SSC task primarily utilize pre-trained models, such as ResNet-50 [36], which are designed for general-purpose tasks and do not provide the specific category-aware features needed for particular segmentation tasks. This limitation hinders the performance of deformable cross-attention mechanisms. **3) Inadequate object perception capability.** Despite progress in geometric integrity, issues such as long traces of objects in scenes and object loss still persist.

In this work, we propose an effective method for achieving precise semantic scene completion using only a single RGB image, which not only reduces the computational burden but also enhances its application potential in specific scenarios. *To address the issue of negative voxel query proposal generation,* we consider weakening the interaction of geometrically relevant but semantically irrelevant image features (e.g., sky) associated with negative query proposals, while simultaneously enhancing the information interaction between positive query proposals and areas of interest. But how do we achieve this? Inspired by the Contrastive Language-Image Pre-Training (CLIP) [37] model, we employ additional textual priors to optimize the distribution of image features. This not only enhances the focus on semantic features relevant to the task but also weakens the influence of those unrelated to it. In addition, it also *addresses the problem of category-agnostic feature extraction*, thereby serving dual purposes. *To address the issue of inadequate object perception capability,* we draw inspiration from UniVS [38] to introduce a set of learnable object queries for better perceiving objects within the scene. Building on this, we iteratively update query proposals to enhance the perception of semantics and objects by using the query-to-query cross-attention, which aggregates information flow from object-to-scene and scene-to-object. Compared to previous methods, a detailed comparison is shown in Table I.

In summary, our contributions of this work are as follows.

- We introduce a straightforward yet effective Semantic-aware Guided (SAG) module, which can be seamlessly integrated with task-related semantic priors to enhance the effective interaction between image features and voxel query proposals in a plug-and-play manner.
- We propose an Interactive Refinement Transformer (IRT) block, which iteratively updates voxel query proposals to enhance the perception of semantics and objects within the scene by using the query-to-query cross-attention, which aggregates information flow from object-to-scene and scene-to-object.
- Extensive experiments demonstrate that our approach outperforms state-of-the-art monocular methods on the SemanticKITTI and SSCBench-KITTI-360 datasets. Additionally, our method also shows superior performance in the generation of small objects.

II. Related Work

In this section, we review the literature concerning the SSC task. We begin with a concise summary of indoor SSC efforts that leverage depth maps or additional RGB images as input. Subsequently, we move on to more complex and challenging outdoor scenes.

A. Semantic Scene Completion for Indoor Scenes

1) Depth-Based Semantic Scene Completion: Semantic scene completion aims to reconstruct the complete 3D structure from limited observations while simultaneously assigning semantic labels to each voxel. As a pioneering work, Song et al. [42] proposed the SSCNet, which processes a single depth image to simultaneously predict voxel occupancy and semantic labels by using an expanded context convolutional module. Moreover, they also employ the proposed flipped Truncated Signed Distance Function (f-TSDF) for encoding, which facilitates strong gradient guidance near the surfaces of objects.

After SSCNet, a wave of research [43], [44], [45], [46], [47], [48] has sprung up, focusing on advancements such as feature extraction [43], [44], network architecture design [45], [46] and the optimization of computational efficiency [47], [48]. For example, Guo and Tong [43] proposed the VV-Net, which utilizes a 2D-CNN to extract geometric features from depth images and then maps them onto the 3D voxel grids for

TABLE I

COMPARISON AMONG SOME REPRESENTATIVE METHODS. "TASK-RE. SEM.-AWARE" DENOTES THE TASK-RELATED SEMANTICS. "NVQ" DENOTES THE NEGATIVE VOXEL QUERY. "IMAGE BACKB." REPRESENTS THE IMAGE BACKBONE. "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

Method	prior-assisted	task-re. sem.-aware	NVQ mitigation	object-aware	image backb.	2D-3D transform	3D representation
MonoScene [CVPR22] [29]	✗	✗	✗	✗	EfficientNetB7	FLoSP-based	Voxel
TPVFormer [CVPR23] [18]	✗	✗	✗	✗	EfficientNetB7	View-TPV	TPV
VoxFormer [CVPR23] [30]	Depth	✗	✗	✗	ResNet-50	3D-2D projection	Voxel
OccFormer [ICCV23] [32]	✗	✗	✗	✗	EfficientNetB7	Lift-splat-based	Voxel
StereoScene [IJCAI24] [39]	✗	✗	✗	✗	EfficientNetB7	View-BEV& stereo	Voxel
Symphonies [CVPR24] [31]	Depth	✗	✗	✓	ResNet-50-MD	3D-2D projection	Voxel
SparseOcc [CVPR24] [40]	✗	✗	✗	✓	EfficientNetB7-FPN	Lift-splat-based	Voxel
HASSC [CVPR24] [41]	Depth	✗	✗	✗	ResNet-50	3D-2D projection	Voxel
Ours	Depth/Text	✓	✓	✓	ResNet-50-MD	3D-2D projection	Voxel

computation. Zhang et al. [45] introduce the CCPNet, which enhances label consistency in pyramid contexts and employs a guided residual refinement module for progressive indoor scene reconstruction. Wang et al. [46] proposed the ForkNet, which utilizes a multi-branch architecture. Specifically, they use a unified encoder for feature extraction and three independent decoders dedicated to predicting incomplete and complete geometric structures alongside semantic volumes. Furthermore, Zhang et al. [47] proposed the spatial group convolution network, which accelerates dense reconstruction tasks by grouping along spatial dimensions. Despite the significant advancements achieved by these methods, current research has not yet fully leveraged the color and texture information present in RGB images.

2) Integrating Depth Map and RGB Image for Semantic Scene Completion: Considering the rich color information contained in RGB images, recent works [49], [50], [51], [52], [53], [54], [55], [56] leverage RGB images as a complement to depth maps so as to enhance the performance of the SSC task. These efforts primarily focus on multimodal data fusion [49], network optimization [50], [51], [52], and depth data transformation [53], [54].

Garbade et al. [49] proposed TS3D, a dual-stream convolutional network for semantic scene completion. This approach first performs semantic segmentation on RGB images using pre-trained ResNet [36] model, then maps the segmentation results to the 3D space and finally employs 3DCNN for context-aware inference of complete semantic scene information. Li et al. [50] proposed DDRNet, a lightweight dimension-decomposition residual network. By incorporating dimension-decomposition residual modules, the network reduces its parameters. Moreover, they use a multi-scale fusion strategy to adapt to objects of varying sizes. Immediately, they proposed AICNet [51], which achieves anisotropic three-dimensional receptive fields by decomposing three-dimensional convolutions into three consecutive one-dimensional convolutions, each with an adaptively sized convolution kernel. Liu et al. [52] presented GRFNet, the first semantic scene completion network utilizing gated recurrent units. Notably, some researchers [54], [55], [56] also attempt to improve depth maps by converting them into a three-channel HHA image (including disparity, height above ground, and the angle between the pixel's local surface normal and the direction of gravity) for feature extraction. Although these methods further enhance the performance of SSC by utilizing additional RGB images, they are predominantly applied to smaller indoor scene datasets.

B. Semantic Scene Completion for Outdoor Scenes

Semantic scene completion methods for outdoor scenes can be classified into three categories according to the form of input: LiDAR-based [57], [58], [60], [61], [62], [63], [64], surrounding view-based [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] and monocular camera-based [29], [30], [31], [32], [33] methods.

1) LiDAR-Based Semantic Scene Completion: Chen et al. [57] proposed S3CNet, a network designed for semantic scene completion using point clouds. Due to the sparsity of LiDAR point clouds, especially in distant areas, direct feature extraction is rather challenging. To address this, they adopt multi-view fusion and semantic post-processing strategies to effectively perceive small-scale objects in distant or occluded regions. Subsequent studies further enhance SSC performance by incorporating point-voxel interaction modules [58], [59], local implicit functions [60] and additional RGB images [61]. Notably, SCPNet [62] leverages a knowledge distillation model to enhance the representational learning capability of the network. Additionally, to address the dynamic trajectory issue arising from the direct fusion of multi-frame point clouds, they use panoramic segmentation labels to correct inaccurate dynamic object labels, thereby improving model accuracy. Zuo et al. proposed PointOcc [63], which transforms point clouds into projected views using the cylindrical Tri-perspective view (TPV) strategy and then extracts features using a pre-trained 2D backbone network. Recently, Cao et al. [64] proposed the panoramic scene completion framework PaSCo that integrates per-voxel and per-instance uncertainties to pave the way for applications in robotics and autonomous driving.

Although these works have advanced the field of semantic scene completion, they depend on costly Lidar sensors. Motivated by the remarkable success of Tesla's Transformer-based purely visual input in autonomous driving, recent research has shifted towards exploring camera-only approaches to broaden the scope of investigation in this domain.

2) Surrounding View-Based Semantic Occupancy Prediction: Driven by the rapid advancement of BEV perception [8],

[65], [66], researchers proposed the task of semantic occupancy prediction. Unlike the traditional definition of SSC, which aims to recover the geometry and semantic information of a complete scene from limited observations, semantic occupancy prediction focuses on representing and understanding 3D space through occupancy grids. Compared to 3D detection tasks that concentrate on coarse-grained object representation, 3D occupancy provides a fine-grained description of the physical world. In the past couple of years, research on surrounding view-based semantic occupancy prediction has flourished [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], with works primarily falling into the following categories: 1) image-voxel projection methods [9], [10], [11], [12], [13], [14], [15], [16], [17], 2) Tri-Perspective View methods [18], [19], [20], [21], and 3) nerf-based self-supervised methods [22], [23], [24], [25], [26], [27], [28]. Wei et al. [12] proposed the innovative work SurroundOcc, in which they use multiple images as input, extract multi-scale features and then elevate these features to 3D space using spatial 2D-3D attention mechanisms. Moreover, they employ 3D convolutions to progressively obtain upsampled features. OctOcc [17] uses a hierarchical octree and a filter mask prediction network to minimize redundancy in voxel queries and image features, respectively. However, it relies on additional projected LiDAR depth information for supervision. Huang et al. [18] proposed the renowned TPVFormer, which, unlike voxel and BEV representations, models each point in three-dimensional space by summing up its projection features on three planes and employs a Transformer-based TPV encoder to acquire TPV features. Recently, Zhang et al. [23] proposed the self-supervised semantic occupancy prediction framework OccNerf. Unlike previous works that rely on dense 3D annotations, this approach only uses 2D segmentation and depth images obtained from existing pre-trained models for supervision, significantly reducing the burden of extensive annotation efforts.

3) Monocular Camera-Based Semantic Scene Completion: Different from the surrounding view-based semantic occupancy prediction, monocular camera-based semantic scene completion works [29], [30], [31], [32], [33] primarily predict the voxel-wise occupancy state and corresponding semantic labels of 3D scenes from a single image. Cao and Charette [29] proposed Monoscene, the first monocular semantic scene completion framework, which employs a FLoSP module to map 2D features into 3D space and then employs a 3D contextual relationship prior module to enhance the extraction of long-range context information. Considering the depth ambiguity inherent in mapping 2D features into 3D space, some studies attempt to assist SSC tasks with offline depth maps [30], [31], [32], [33] and multimodal information [33]. Notably, VoxFormer [30] utilizes depth maps generated by a pre-trained model to guide its voxel occupancy in 3D space and then adopts the deformable cross-attention mechanism [35] to facilitate interaction between 3D query proposals and image features. Recently, Jiang et al. [31] proposed Symphonies, which enhances SSC performance from the perspectives of instance-awareness and holistic context understanding through stacked cross-attention modules. Although the aforementioned works use only a single image as input, thereby reducing the computational requirements for processing surrounding views, they neglect the difference between task-relevant semantic information and general feature information, leading to sub-optimal results. Specifically, these studies mostly utilize pre-trained image feature extraction models, such as ResNet-50 [36], EfficientNetB7 [67] and Transformer [68], for feature extraction. However, these general models are not specifically designed for the semantic scene completion task, leading to the neglect of specific semantic category information or the extraction of unnecessary semantic category information. To address this, we introduce a straightforward yet effective Semantic-aware Guided (SAG) module that seamlessly integrates with task-related semantic priors in order to facilitate effective interactions between image features and voxel query proposals in a plug-and-play manner.

III. Method

In this work, we propose a monocular semantic scene completion method that employs the Semantic-aware Guided (SAG) module and Interactive Refinement Transformer (IRT) block to estimate the geometric occupancy and per-voxel semantic labels of 3D scenes from a single RGB image. The overall architecture of our method is illustrated in Fig. 2, which can be divided into five steps:

- Utilize existing fundamental models to extract multi-scale features, depth maps, textual tags and captions from RGB images.
- Back-project the depth map into a point cloud, then voxelize it to obtain the voxel query proposals.
- Integrate task-related semantic priors into the image features to facilitate effective interactions between image features and voxel query proposals.
- Employ the IRT block to iteratively update voxel queries, which enhance the perception of semantics and objects within the scene by using the query-to-query cross-attention.
- Predict the dense semantic map using the predefined segmentation head.

A. Semantic-Aware Guided Module

We first revisit the general workflow of existing methods for the information aggregation between voxel query proposals $Q_p \in R^{N_p \times d}$ and image features $F_{img} \in \mathbb{R}^{b \times h \times w}$. Then, we analyze the inherent issues in this process and introduce this proposed plug-and-play semantic-aware guided module.

1) Feature Extraction: Leverage common image feature extractors such as ResNet-50 [36] or EfficientNetB7 [67] to extract features from a single image, yielding image features $F_{img} \in \mathbb{R}^{b \times h \times w}$.

2) Predefine Voxel Queries: Predefine the range of voxel queries $Q \in \mathbb{R}^{N_v \times d}$ ($N_v = h \times w \times z$), which comprises a set of learnable parameters in a 3D grid shape.

3) Binary Voxel Query Proposals: Convert depth maps I_d, generated by off-the-shelf depth estimators such as Mobilestereonet [69] or DiffusionDepth [33], into point clouds, which are then voxelized. Subsequently, each voxel is

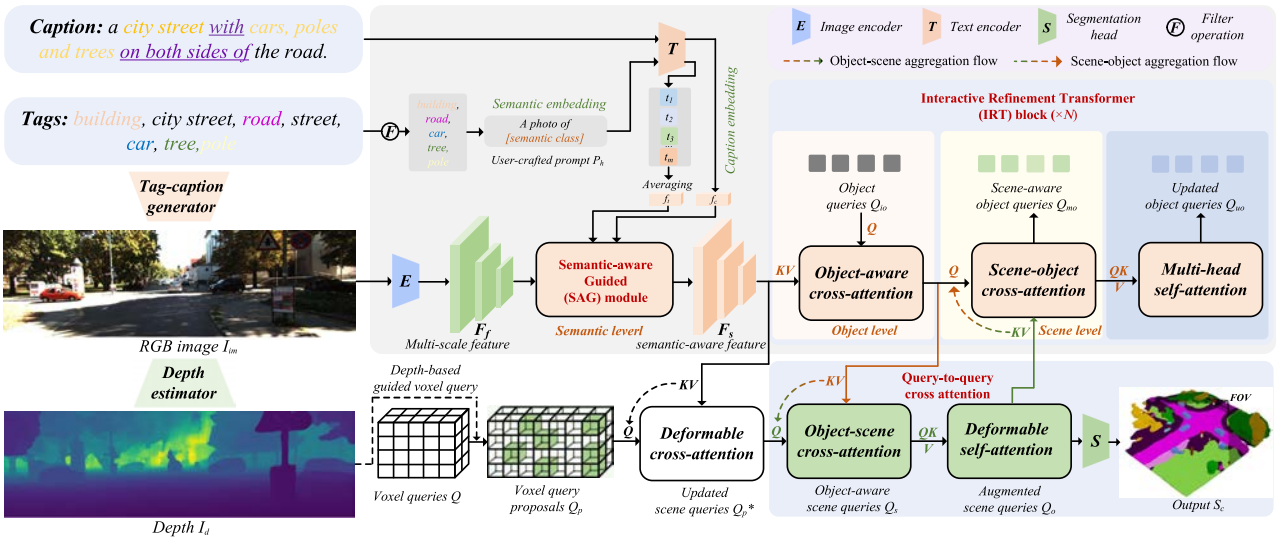

Fig. 2. Overall framework of our method. Given a single image, we first utilize existing foundational models to obtain corresponding multi-scale features, a depth map, tags and captions. Then, we apply a filtering operation to extract task-related tags and integrate them with captions into the image features, resulting in semantic-aware features via the proposed plug-and-play Semantic-aware Guided (SAG) module. Subsequently, we introduce a set of learnable object queries and further propose an Interactive Refinement Transformer (IRT) block to iteratively update query proposals by using the query-to-query cross-attention. Here, this query-to-query cross-attention integrates object-scene and scene-object aggregation flows, thus enhancing the perception of semantics and objects within the scene. Other details in the IRT block, such as feed-forward networks and normalization layers, are omitted for a neat presentation.

set to 1 if it is occupied. This process yields binary occupancy query proposals $Q_p \in \mathbb{R}^{N_p \times d}$ $(N_p \leq h \times w \times z)$ derived from Q.

4) Deformable Feature Aggregation: Utilize the deformable cross-attention mechanism [35] to dynamically determine the key regions on the image plane F_{img} for each voxel query feature, thereby optimizing Q_p. Specifically, we first compute reference points R_i on the image plane for each voxel query feature center $q_i = (x, y, z)$ via the camera projection matrix $T_{\text{cam-world}}$.

$$R_i = [R, C] \cdot T_{\text{cam-world}} \cdot q_i, \tag{1}$$

where R and C represent the rotation matrix and the calibration matrix, respectively. $[\cdot]$ denotes the combination function. Then, the deformable cross-attention feature aggregation can be calculated as follows.

$$\begin{aligned}&\text{DeformCrossAtten}\left(q_i, R_i, F_{img}\right)\\&= \sum_{s=1}^{S} \mathbf{W}_s \left[\sum_{k=1}^{K} A_{sqk}\left(q_i\right) \cdot \mathbf{W}'_s F_{img}\left(R_i + \Delta R_{sqk}\right)\right].\end{aligned} \tag{2}$$

Herein, S denotes the number of split attention heads. $\mathbf{W}_s$ and $\mathbf{W}'_s$ represent learnable weights. A_{sqk} denotes the MLP for aggregating attention scores. The variable K indicates the number of sampling positions, while ΔR_{sqk} represents the sampling offset.

Although the aforementioned information aggregation process avoids considering the spatial positions of entire image pixels and focuses only on the most relevant positional information near the geometry, we realize that there are still two key issues: 1) Common image feature extractors are primarily designed for conventional tasks. However, for the specific semantic segmentation task, they fail to accurately perceive the regions of interest for specific categories, thereby limiting the performance of deformable cross-attention. 2) The quality of voxel query proposals is highly dependent on the accuracy of depth estimators. However, the scale ambiguity and occlusion may lead to inaccuracies in depth estimation, thereby affecting the accuracy of voxel query proposals and hindering effective 3D-2D correspondence and information aggregation.

To address the aforementioned issues, we introduce a straightforward yet effective semantic-aware guided (SAG) module, which seamlessly integrates with task-related semantic priors to facilitate effective interactions between image features and voxel query proposals in a plug-and-play manner. Unlike OctOcc [17], which relies on projected LiDAR depth information processed through binarization and dilation operations to generate dilated depth masks that filter out irrelevant image regions, our approach utilizes additional textual priors extracted from pre-trained models to optimize the distribution of image features. This not only enhances the focus on semantic features relevant to the task but also weakens the influence of those unrelated to it. Importantly, our method does not rely on additional LiDAR depth information for supervision during training. To put it succinctly, SAG is designed to enhance the perception of specific semantic regions and weaken the interaction with geometrically relevant but semantically irrelevant features associated with negative query proposals.

We first utilize existing foundational models to extract multi-scale features, textual tags and image captions. Specifically, we employ ResNet-50 [36] as the backbone and leverage the pre-trained weights from the MaskDINO [70] model to extract multi-scale features $\{F_f\}_{s=1}^{s} \in \mathbb{R}^{b \times h_s \times w_s}$ from a single RGB image I_{im}, where s denotes the scale of the s-th downsampled feature map. Next, we utilize the pre-trained Tag2Text model [71], which is based on BLIP [72], to generate tags along with corresponding image captions. Notably, we use the pre-trained clip text model [37] to filter out tags that are not

present in the target datasets by setting a similarity threshold. This ensures that we focus only on the semantic categories pertinent to our task.

Next, we construct textual prompts from semantic categories as *"A photo of [semantic class]"*, where *[semantic class]* denotes the category name. The reason for this design is to align closely with the text input format used during the original training of the CLIP model. Following the method [38], we input each textual prompt into a tokenizer to generate string tokens, which are subsequently fed into the text encoder of CLIP. By performing an averaging operation, we obtain a comprehensive feature representation f_t that encompasses specific semantic categories. Similarly, we input the generated image captions into the CLIP text encoder to obtain caption features. Compared to tags that only include categories, captions also incorporate spatial relationships, such as expressions like *"with"* and *"on both sides of"*. Here, we denote the caption features as f_c.

Then, we introduce two consecutive cross-attention layers to enable interactions between semantic category embeddings and flattened multi-scale image embeddings, as well as between image captions and image embeddings. We first present the interaction between semantic category embeddings and flattened multi-scale image embeddings. We use Multi-Layer Perceptrons (MLPs) to perform dimensional projection on both semantic category embeddings and flattened image embedding features, yielding projected features f_{tl} and F_{fl}. Herein, the query is the projected semantic category features, while the keys and values are the projected multi-scale image embedding features. The cross-attention interaction process can be succinctly defined as follows.

$$F_f^* = \text{TI-CrossAtten}\left(f_{tl}, F_{fl}\right). \tag{3}$$

Different from the original cross-attention mechanism, we incorporate a temperature adaptive coefficient into the softmax function to further adjust the smoothness of the softmax output. Consequently, the temperature adaptive softmax (Ta-softmax) function in TI-CrossAtten is defined as follows.

$$a_{tl} = \text{Ta-softmax}\left(\frac{S}{\frac{1}{\epsilon}\log\left(\frac{\delta(C-1)+1}{1-\delta}\right)}\right), \tag{4}$$

where ϵ and δ represent pre-set parameters, which are set to 0.1 and 0.5, respectively. C represents the number of features. S represents the attention score matrix. Similarly, we implement the interaction between the caption embeddings and image embeddings to further perceive the spatial distribution. We define this interaction as follows. *For simplify, we omit dimension transformation operations here.*

$$F_f' = \text{CI-CrossAtten}\left(f_{cl}, F_f^*\right). \tag{5}$$

Finally, we obtain image features enhanced by semantic-aware guidance via an element-wise addition operation.

$$F_s = \lambda F_f + (1-\lambda)F_f', \tag{6}$$

where λ represents a hyper-parameter, which is set to 0.15.

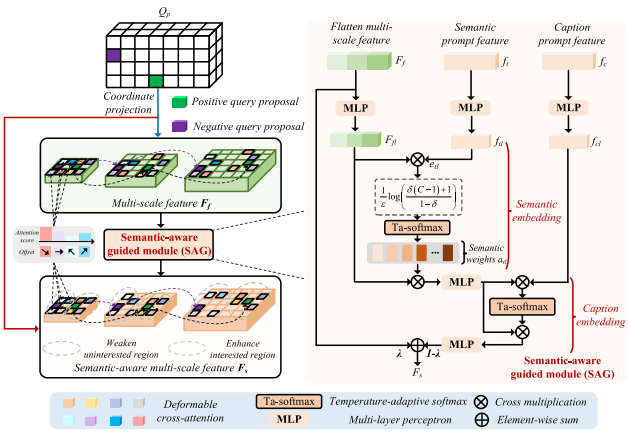

Fig. 3. Illustration of the Semantic-aware Guided (SAG) module, which can be directly integrated into existing frameworks to facilitate effective interaction between voxel query proposals and image features.

B. Interactive Refinement Transformer Block

Following VoxFormer [30], we convert estimated depth maps into point clouds, which are then voxelized and subjected to binary occupancy processing to generate voxel query proposals Q_p from predefined voxel queries Q. Unlike VoxFormer, we remove the U-Net network previously used for depth correction. This modification is made because the SAG module weakens the interaction with geometrically relevant but semantically irrelevant features associated with negative query proposals. Next, we detail our proposed Interactive Refinement Transformer (IRT) block, which iteratively updates query proposals to enhance the perception of semantics and objects within scenes by integrating the object-scene aggregation flow and scene-object aggregation flow. *Due to the detailed exposition of the deformable feature aggregation process and formulas in the previous section, we omit the details in the following formulas to avoid redundancy.*

To better perceive objects within scenes, inspired by [40], [73], and [74], we define a set of learnable object queries $Q_{io} \in R^{N_o \times d_o}$ and iteratively update them in the subsequent process, where N_o represents the number of queries and d_o denotes their dimensions. Therefore, we define the information aggregation between the object queries Q_{io} and image features F_s as follows.

$$Q_{io}^* = \text{ObjectAware-DeformCrossAttn}\left(Q_{io}, R_{io}, F_s\right), \tag{7}$$

where R_{io} represents the learnable 2D reference points.

Next, we introduce the information aggregation flow from object to scene. In this process, image features are first defined as keys and values, and Qp serves as queries. By employing the deformable cross-attention, we efficiently map image features onto relevant voxel query positions, obtaining updated voxel queries Qp^*. Subsequently, we utilize a cross-attention to effectively integrate object queries Q_{io}^* into scene queries Q_p^* and obtain object-aware scene queries Q_s. To better perceive the entire scene's information, we employ a deformable self-attention to aggregate contextual information, resulting in updated scene queries Q_o.

$$Q_p^* = \text{DeformCrossAttn}\left(Q_p, R_p, F_s\right), \tag{8}$$

$$Q_s = \text{CrossAttn}\left(Q_p^*, Q_{io}^*, Q_{io}^*\right), \tag{9}$$

$$Q_o = \text{DeformSelfAttn}\left(Q_s, Q_s\right), \tag{10}$$

where R_p represents the reference points on the image plane for the query proposal Q_p through coordinate projection transformation.

Then, we introduce the information aggregation flow from scene to object. In this process, we use deformable cross-attention to update object queries Q_{io}^* with updated voxel query proposals Q_o, achieving scene-aware object queries Q_{mo}. Subsequently, by incorporating a multi-head self-attention, we dynamically adjust the focus and attention range of the objects, resulting in updated object queries denoted as Q_{uo}.

$$Q_{mo} = \text{DeformCrossAttn}\left(Q_{io}^*, R_{io}^* \cdot T_{2D-3D}, Q_o\right), \tag{11}$$

$$Q_{uo} = \text{MultiHead-SelfAttn}\left(Q_{mo}, Q_{mo}\right), \tag{12}$$

where R_{io}^* represents the 2D reference points and T_{2D-3D} denotes the 2D-3D projection transformation.

Finally, inspired by [29], our segmentation head, which consists of a series of 3D convolutions and ASPP convolutions [75], outputs semantic scene completion results. We employ the scene-class affinity loss L_{scal} [29] for model training, which simultaneously considers both geometric and semantic supervision. Additionally, following [30], we incorporate a class-frequency weighted cross-entropy loss L_{ce} to enhance model performance.

$$L = L_{scal}^{geo} + L_{scal}^{sem} + L_{ce}. \tag{13}$$

Here, L_{scal}^{geo} and L_{scal}^{sem} represent the optimization of geometric and semantic aspects, respectively. Furthermore, following the DETR series [35], we also apply auxiliary loss [31] for enhanced supervision after each IRT block layer, which is scaled by a factor of 0.5. We denote the semantic scene completion results as S_c.

IV. Experiments

In this section, we first introduce the datasets used in our experiments, followed by the implementation details and evaluation metrics. We then compare the performance of our proposed method with other competitive algorithms on widely used datasets. Finally, we conduct a series of ablation studies to validate the effectiveness of our method's module designs, accompanied by comparative statistical analyses on model parameters.

A. Datasets

1) The SemanticKITTI Dataset: The SemanticKITTI dataset, designed for semantic scene understanding, was introduced at the International Conference on Computer Vision (ICCV) in 2019 [7]. As the first benchmark dataset for outdoor semantic scene completion, it includes LIDAR and RGB image data from 22 sequences. In the evaluation setup of semantic scene completion tasks, this dataset merges multiple LIDAR point cloud frames as ground truth based on the semantic labels of point clouds. Specifically, it constructs a voxel grid of size $256 \times 256 \times 32$ with a voxel resolution of $0.2m$, which is annotated with 19 semantic category labels and one empty label. Following the data partitioning strategy of prior studies [29], [30], sequences 0-7 and 9-10 serve as the training set, sequence 08 as the validation set and sequences 11-22 as the test set. In accordance with [30], RGB images captured by the left camera are used as monocular input.

2) The SSCBench-KITTI-360 Dataset: The SSCBench-KITTI-360 dataset, introduced in 2023 [76], aims to address the issues of long traces and limited geographical coverage presented in the SemanticKITTI dataset. SSCBench-KITTI-360 employs 3D bounding boxes to effectively address traces of dynamic objects when aggregating multi-frame point clouds. It includes 9 sequences, with sequences 00, 02-05, 07 and 10 designated as the training set, sequence 06 as the validation set and sequence 09 as the test set. According to the [7] configuration, the voxel grid dimensions are $256 \times 256 \times 32$ with a resolution of $0.2m$, extending $51.2m$ forward, $25.6m$ to each side and $6.4m$ in height. This dataset provides a more comprehensive benchmark for the semantic scene perception analysis of complex dynamic environments.

B. Implementation Details

We implement our proposed method using the PyTorch framework. The image encoder utilizes ResNet-50 [36] as the backbone, initialized with pre-trained MaskDINO [70] weights. For depth map extraction, we employ the off-the-shelf pre-trained Mobilestereonet [69] model. Additionally, we follow the Bridge3D [77] method to extract tags and captions using the pre-trained Tag2Text [64] model. In our experiments, we set the initial number of object queries to 100. We adhere to the same evaluation settings as in [29] and [76] to ensure fair comparisons on the SemanticKITTI and SSCBench-KITTI-360 datasets. For training, we use four NVIDIA 3090 GPUs with a batch size of 4. We use the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. Our model are trained for 26 epochs.

C. Evaluation Metrics

Following previous methods [29], [30], in our experiments, we evaluate both class-agnostic geometric completion and semantically aware scene completion. For geometric completion, we use the Intersection over Union (IoU) as the evaluation metric. For semantically aware scene completion, we employ the mean Intersection over Union (mIoU) to assess the performance. These metrics allow us to thoroughly evaluate and compare the effectiveness of different approaches at both the geometric and semantic levels of scene completion.

D. Semantic Scene Completion Results

1) Results on SemanticKITTI Validation Set: We compare our proposed method with ten classical approaches, including three recent papers [31], [40], [41] published at Conference on Computer Vision and Pattern Recognition (CVPR) 2024, on the SemanticKITTI validation set. Table II demonstrates that our method achieves comparable performance to these newly presented methods, with notably superior semantic

TABLE II

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE OUTDOOR SCENES FROM THE SEMANTICKITTI VALIDATION SET, IN TERMS OF IOU AND MIOU METRICS. † REPRESENTS THE RESULT OBTAINED USING ONLY A SINGLE IMAGE. * REPRESENTS THE RESULT REPORTED FROM OCCGEN [80]. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN **BOLD**

Method	SSC Input	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
LMSCNet [3DV20] [78]	x_{3D}^{occ}	28.61	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00	6.70
3DSketch [CVPR20] [79]	x^{rgb}, x^{TSDF}	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] [51]	x^{rgb}, x^{depth}	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] [29]	x^{rgb}	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] [18]	x^{rgb}	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] [30]	x^{rgb}	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
OccFormer [ICCV23] [32]	x^{rgb}	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
Symphonies* [CVPR24] [31]	x^{rgb}	41.44	55.78	26.77	14.57	0.19	18.76	27.23	15.99	1.44	2.28	9.52	24.50	4.32	28.49	3.19	8.09	0.00	6.18	**8.99**	**5.39**	13.44
SparseOcc [CVPR24] [40]	x^{rgb}	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
HASSC-S† [CVPR24] [41]	x^{rgb}	**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
Ours	x^{rgb}	40.52	59.34	27.87	16.21	**1.10**	15.85	**27.69**	16.07	**1.50**	**2.32**	**9.61**	21.01	**6.58**	28.65	**3.72**	**8.64**	0.00	**7.78**	5.13	2.79	**13.78**

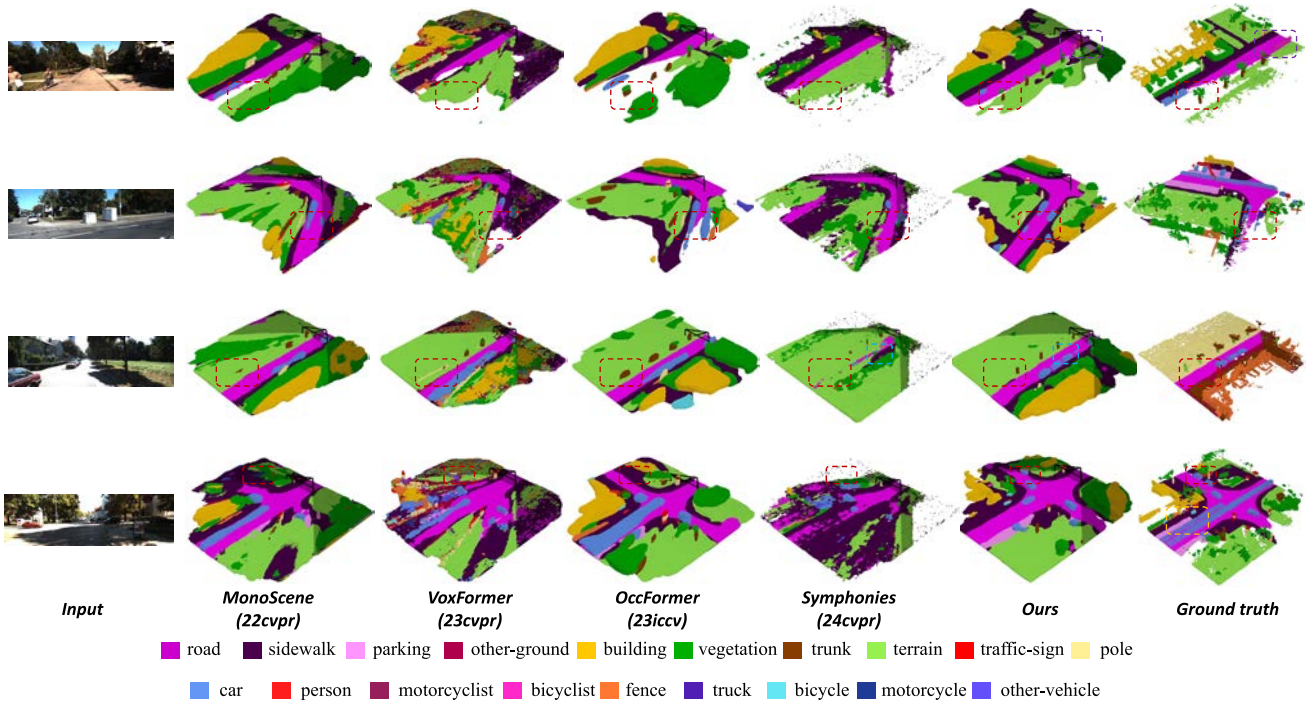

Fig. 4. Visual comparison of our method with state-of-the-art methods on the SemanticKITTI validation set. Compared to other competitive methods, our method generates more rational semantic spatial layouts, fewer semantic confusion overlap points and clearer object boundaries. Best viewed in colors.

mIoU scores. These results are derived from their original papers, with the Symphonies results sourced from the recent OccGen [80]. Compared to the latest method, HASSC-S, our approach shows an improvement of 0.30 in mIoU. Although our method does not excel in categories like road, sidewalk and parking, it outperforms previous methods in categories such as car, bicycle, motorcycle, truck and bicyclist. Compared to classical methods OccFormer and Symphonies, our method achieves overall improvements of 0.34 and 0.32, respectively. In the categories of car, bicycle, motorcycle, and person, our method shows improvements of 2.6, 0.69, 1.13, and 0.94 over OccFormer, and 0.46, 0.06, 0.04, and 0.53 over Symphonies, respectively. Although these categories account for a low proportion in the dataset (for example, persons represent only 0.07%), their improvements in mIoU are crucial for safe autonomous driving. Additionally, we observe that almost all methods struggle in the motorcyclist category, which further motivates us to enhance our method's perception capabilities under conditions of limited training samples and the rapid movement of small objects.

We also perform a qualitative comparison of our proposed method with Monoscene, VoxFormer, OccFormer and Symphonies, as shown in Fig. 4. It is important to note that the results for VoxFormer are obtained using only a single RGB image as input. Intuitively, we observe that both our method and Symphonies exhibit no significant long traces in the car category perception and perform better than other methods (highlighted with red dashed lines). We argue that

TABLE III

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE OUTDOOR SCENES FROM THE SEMANTICKITTI TEST SET, IN TERMS OF IOU AND MIOU METRICS. † REPRESENTS THE RESULT OBTAINED USING ONLY A SINGLE IMAGE. * REPRESENTS THE RESULT REPORTED FROM [36].THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN **BOLD**

Method	SSC Input	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
LMSCNet [3DV20] [78]	$\hat{x}^{occ}_{3D}$	31.38	46.70	19.50	13.50	3.10	10.30	14.30	0.30	0.00	0.00	0.00	10.80	0.00	10.40	0.00	0.00	0.00	5.40	0.00	0.00	7.70
3DSketch [CVPR20] [79]	x^{rgb},$\hat{x}^{TSDF}$	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] [51]	x^{rgb},$\hat{x}^{depth}$	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] [29]	x^{rgb}	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] [18]	x^{rgb}	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] [30]	x^{rgb}	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
OccFormer [ICCV23] [32]	x^{rgb}	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
Symphonies* [CVPR24] [31]	x^{rgb}	34.53	55.70	26.80	25.30	4.90	**21.30**	22.10	1.90	**1.70**	1.30	5.80	22.90	8.20	19.50	**2.20**	1.30	**0.50**	13.10	**6.80**	**5.80**	13.02
HASSC†-S [CVPR24] [41]	x^{rgb}	43.40	54.60	27.70	23.80	6.20	21.10	22.80	**4.70**	1.60	1.00	3.90	**23.80**	8.50	**23.30**	1.60	4.00	0.30	13.10	5.80	5.50	13.34
Ours	x^{rgb}	38.44	**56.20**	30.20	28.40	**6.80**	18.50	**22.90**	**4.70**	**1.70**	**1.90**	**5.90**	20.10	**8.60**	19.10	**2.20**	**4.10**	**0.50**	**13.30**	4.20	**5.80**	**13.42**

this primarily results from two factors: *1) the labels provided by the SemanticKITTI dataset are somewhat inaccurate, as evidenced by the car traces issue (indicated by yellow dashed lines), 2) both our method and Symphonies incorporate learnable object queries to more accurately detect and perceive object instances. However, Symphonies occasionally misclassifies cars as vegetation (as shown in the third row with blue dashed lines). We argue that for a complete scene representation, accurately identifying the number of objects and reconstructing reasonable geometric shapes is crucial. This not only aids in more precise future planning but also significantly enhances the perception of the entire scene.* Additionally, we note that both VoxFormer and Symphonies exhibit semantic confusion in processing vegetation, building and road categories. In contrast, our method effectively avoids the aforementioned issues. Last, but equally important, our approach can perceive the presence of bicyclists, as shown in the first row with purple dashed lines, which is crucial for the safety of autonomous driving.

2) Results on SemanticKITTI Test Set: We compare our method against nine other methods on the SemanticKITTI test set. It is important to note that the test set does not provide ground-truth labels for the corresponding scenes. Additionally, these results are cited from their original articles and a recent paper [36]. According to data presented in Table III, our method achieves a mean Intersection over Union (mIoU) improvement of 0.08 compared to the latest method HASSC-S [41]. Although the overall increase in performance is modest, we observe significant improvements in the categories of cars, motorcycles and persons, with mIoU increases of 0.10, 0.90 and 0.60, respectively. This highlights the superiority of our method in perceiving small objects within complex scenes. We also conduct detailed quantitative comparisons of our method with OccFormer and Symphonies in the categories of car, motorcycle, and bicycle. As demonstrated in Table III, our method exhibits performance enhancements over OccFormer with improvements of 1.3, 0.20, and 3.0, in the categories of cars, bicycles, and motorcycles, respectively. Similarly, when compared to Symphonies, our method achieves performance gains of 0.8, 0.60, and 2.80 in these respective categories. We observe that these improvements align with the expectation that our method continues to exhibit significant advantages on the test set, which features a more diverse range of scenes, compared to Symphonies. We attribute this primarily to the effectiveness of the SAG module, which enables our method to maintain superior generalizability on the test set.

We also conduct a qualitative comparison of our method with Monoscene, VoxFormer, OccFormer and Symphonies on the test set. Specifically, we visualize partial results from sequences 12, 16 and 18. From the first row of Fig. 5 (highlighted with red dashed boxes), it is evident that VoxFormer and Symphonies erroneously generate car objects and incorrect semantic objects, respectively, in scenes lacking the car category. Then, as indicated in the second row of Fig. 5 (highlighted with red dashed boxes), Voxformer, Occformer and Symphonies fail to perceive distant car objects. Additionally, we further observe that our method possesses significant advantages in generating the three-dimensional spatial structure of buildings. For example, Symphonies incorrectly generates buildings as vegetation, as shown in the third row of Fig. 5 (highlighted with red dashed boxes). *We attribute this to the effectiveness of our proposed semantic-aware guided module in conveying aggregated semantic information from the image level to 3D space.* Lastly, it is noteworthy that our method accurately generates distant buildings in the images, a task failed by other methods as depicted in the fourth row of Fig. 5 (highlighted with red dashed boxes). Encouragingly, our method accurately perceives the spatial relationship between vegetation and buildings in the images, which are not directly adjacent in spatial layout, as presented by the yellow dashed boxes in the fourth row of Fig. 5, highlighting the superiority of our method in spatial layout perception.

3) Results on SSCBench-KITTI-360 Dataset Across Different Scales: We compare our proposed method with state-of-the-art approaches, including Monoscene, VoxFormer,

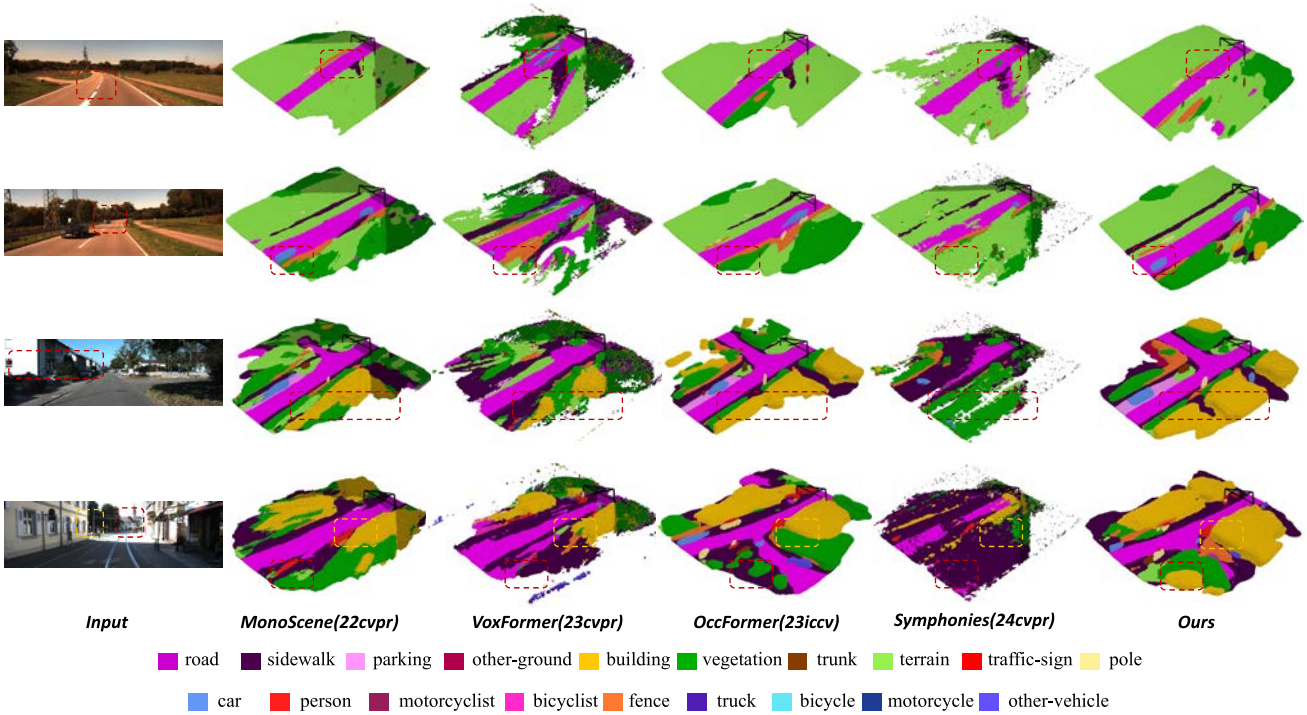

Fig. 5. Visual comparison of our method with state-of-the-art methods on the SemanticKITTI test set. Compared to other competitive methods, our method generates results more consistent with images. Best viewed in colors.

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE OUTDOOR SCENES FROM THE SSCBENCH-KITTI-360 VALIDATION SET AT DIFFERENT SCALES, 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	Scale	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] [29]	12.8	54.65	68.60	44.43	24.32	5.76	45.40	30.83	14.83	1.94	3.25	6.08	42.98	31.96	2.06	9.79	9.28	8.58	9.18	5.86	20.29
	25.6	44.70	59.93	36.05	16.40	4.82	40.60	26.35	12.18	0.83	1.30	4.30	32.75	21.63	1.26	5.91	8.45	7.67	6.76	4.49	16.18
	51.2	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] [30]	12.8	55.45	66.10	38.00	18.44	4.49	41.12	29.41	6.08	2.73	1.97	3.71	45.68	24.70	2.86	8.99	8.84	9.15	10.31	4.40	18.17
	25.6	46.36	58.58	33.63	13.52	4.04	38.24	25.08	6.63	1.73	1.47	3.56	35.16	18.53	2.20	7.43	8.16	9.02	7.02	3.27	15.40
	51.2	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] [18]	12.8	56.56	73.31	48.06	28.38	**6.07**	50.45	36.70	17.72	3.40	5.17	5.78	45.49	31.25	3.60	9.40	11.82	11.17	11.72	5.48	22.50
	25.6	46.83	65.73	39.79	18.35	**5.28**	43.11	30.60	12.94	2.07	2.51	4.74	35.73	22.02	3.19	7.68	9.74	8.49	8.98	3.59	18.03
	51.2	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] [32]	12.8	58.71	73.34	49.76	30.91	5.62	53.65	40.87	22.40	1.94	1.03	8.48	49.91	34.63	4.54	10.64	12.93	14.25	13.81	8.96	23.04
	25.6	47.96	66.53	41.30	20.25	5.45	44.86	33.10	15.21	1.04	0.43	5.21	37.96	24.99	3.79	7.85	10.25	12.37	11.04	6.71	18.38
	51.2	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
Ours	12.8	**60.30**	**76.28**	**50.32**	**32.83**	5.84	**55.49**	**42.04**	**23.36**	**4.26**	**6.73**	**9.05**	**51.64**	**36.12**	**5.24**	**11.71**	**13.83**	**16.75**	**14.59**	**10.04**	**25.89**
	25.6	**52.11**	**70.73**	**44.25**	**23.42**	5.17	**48.95**	**35.20**	**18.69**	**3.41**	**4.07**	**6.64**	**40.31**	**27.86**	**4.13**	**8.40**	**12.08**	**15.14**	**13.27**	**8.11**	**21.66**
	51.2	**43.24**	**58.12**	**35.89**	**15.49**	**4.04**	**41.52**	**24.58**	**11.74**	**2.32**	**2.47**	**4.13**	**33.25**	**20.96**	**3.44**	**6.39**	**9.63**	**10.01**	**8.64**	**5.19**	**16.55**

TPVFormer and OccFormer, on the SSCBench-KITTI-360 validation set. Following VoxFormer's configuration [30], we evaluate across three different ranges in front of the car: $12.8m \times 12.8m \times 6.4m$, $25.6m \times 25.6m \times 6.4m$ and $51.2m \times 51.2m \times 6.4m$. Table IV shows that our method achieves competitive IoU and mIoU at all three scales. Specifically, compared to OccFormer, our method registers mIoU improvements of 2.85, 3.28, and 2.74, and IoU improvements of 1.59, 4.15, and 2.97, respectively. *Compared to the SemanticKITTI dataset, our method exhibits a notably superior performance improvement on the SSCBench-KITTI-360 dataset compared to OccFormer. This advancement is primarily attributed to the more accurate labels provided by SSCBench-KITTI-360, which bolster the effectiveness of the proposed approach.* Furthermore, we observe consistent improvements in categories such as car, bicycle and person, which further substantiates the efficacy of our approach. Specifically, compared to the competitive method OccFormer, our method achieves performance improvements of 1.17, 2.32, and 0.70 in the categories of car, bicycle, and person, respectively, at a scale of 12.8.

TABLE V
ABLATION STUDY ON THE EFFECTIVENESS OF THE SAG MODULE

method	IoU	mIoU
w.o. SAG	39.43	13.21
SAG (w.o. *cap.embedding*)	40.37	13.66
SAG (w. *cap.embedding*)	**40.52**	**13.78**

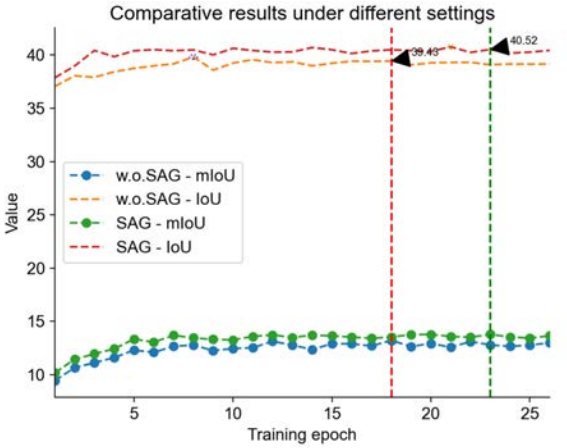

Fig. 6. Analysis of model training curves under different configuration settings. Best viewed in colors.

TABLE VI
ABLATION STUDY ON THE IRT BLOCK

O-S aggregation (8)	O-S aggregation (7)(9)	O-S aggregation (10)	S-O aggregation (11)	S-O aggregation (12)	IoU	mIoU
✓					39.91	13.28
✓	✓				40.29	13.43
✓	✓	✓			40.31	13.46
✓	✓	✓	✓		40.44	13.72
✓	✓	✓	✓	✓	**40.52**	**13.78**

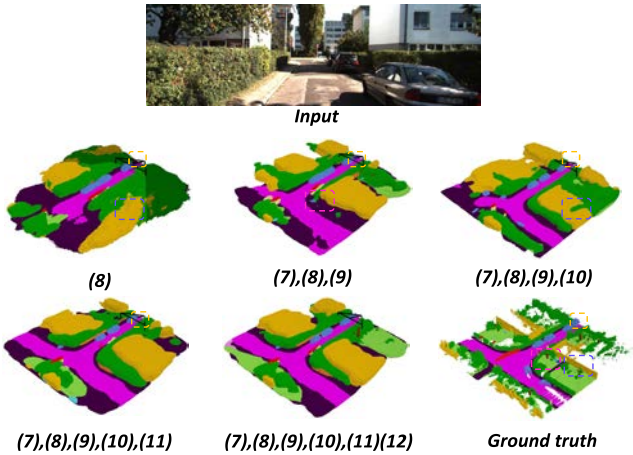

Fig. 7. Ablation study on the effectiveness of the formula settings in the IRT block. Best viewed in colors.

At scales of 25.6 and 51.2, the improvements are 2.10, 2.37, 0.34 and 2.00, 1.66, 0.67, respectively. These enhancements further highlight the advantages of incorporating object queries into our method and demonstrate the IRT block's ability to perceive crucial objects at various scales.

E. Ablation Studies

1) Effectiveness of the Semantic-Aware Guided Module: In this section, we assess the effectiveness of the Semantic-Aware Guidance (SAG) module. Given its plug-and-play nature, we demonstrate its impact by comparing network performance both with and without this module. According to the experimental data in Table V, integrating the SAG module significantly enhances performance, resulting in an increase of 1.09 in IoU and 0.57 in mIoU. We also present the variation in IoU and mIoU metrics on the validation set during training with and without the integration of the SAG module. Fig. 6 shows that incorporating the SAG module not only smoothens the curves but also yields a numerical improvement. Additionally, we explore the necessity of incorporating caption embeddings within the SAG module. The results indicate that including caption embeddings effectively improves the model's perception of various semantic concepts within the scene, leading to increases of 0.15 in IoU and 0.12 in mIoU.

2) Effectiveness of the Interactive Refinement Transformer Block: In this section, we evaluate the effectiveness of the Interactive Refinement Transformer (IRT) block. First, we outline the Object-Scene (OS) information aggregation flow with three different configurations: 1) Eq. (8) serves as a baseline, depicting the interaction between features extracted from images and voxel query proposals; 2) Eq. (7) and (9) are employed to aggregate learnable object queries, further capturing object information within the scene; 3) Eq. (10) leverages self-attention mechanisms to focus on contextual information, thereby comprehensively perceiving the scene's geometric integrity. Subsequently, we introduce the Scene-Object (S-O) information aggregation flow, set up in two forms: 1) Eq. (11) aims to transfer the voxel queries to image-level object queries; 2) Eq. (12) updates object queries through self-attention mechanisms to focus on contextual information. According to the data in Table VI, our method achieves optimal results when integrating the bidirectional information flow, showing an improvement of 0.61 in IoU and 0.50 in mIoU compared to the baseline established by Eq. (8).

Furthermore, we also perform a visual analysis of the model's performance across various formula configurations. All experiments are conducted under the SAG configuration, with adjustments made only to the formula settings within the IRT block. Fig. 7 shows that by integrating all formulas, the model can precisely reconstruct the spatial layout of real scenes, effectively complete missing cars (indicated by yellow dashed boxes), reduce semantic confusion (depicted by purple dashed boxes) and accurately build the local semantic details (highlighted by pink dashed boxes).

3) Effectiveness of the Model Parameter Setting: In this section, we study the impact of model parameter selection on method performance. And we show that optimal results are achieved when parameters ϵ and δ are set to 0.1 and 0.5, respectively, as shown in Table VII. This setting is used by default in our experiments. We also conduct ablation studies on parameter λ. As indicated in Table VII, we test different values of λ at 0.05, 0.10, 0.15, and 0.20. The best results are obtained when λ is set to 0.15; further increasing λ leads to a slight decline in performance.

4) Effectiveness of the Backbone Architecture: In this section, we study the impact of different backbone

TABLE VII

ABLATION STUDY ON THE MODEL PARAMETER SETTING. THE BEST RESULTS ARE IN **BOLD**

Ablation on parameters ε and δ			Ablation on the parameter λ		
ε and δ	IoU	mIoU	λ	IoU	mIoU
ε=0.1, δ=0.3	40.49	13.75	0.05	40.44	13.70
ε=0.1, δ=0.5	**40.52**	**13.78**	0.10	40.48	13.74
ε=0.1, δ=0.7	40.50	13.76	0.15	**40.52**	**13.78**
ε=0.1, δ=0.9	40.48	13.75	0.20	40.50	13.77

TABLE VIII

ABLATION STUDY ON THE IMPACT OF DIFFERENT BACKBONE ARCHITECTURES ON RESULTS. THE BEST RESULTS ARE IN **BOLD**

method	ResNet-101-DCN	ResNet-50-MD	SwinL-DINO	SwinL-MD
IoU	39.79	40.52	41.51	**41.83**
mIoU	13.72	13.78	13.76	**13.91**

TABLE IX

COMPLEXITY ANALYSIS ON THE SEMANTICKITTI VALIDATION SET. WE REPORT THE PARAMS, FLOPS, LATENCY, IOU AND MIOU. THE BEST METRIC WITHIN EACH COLUMN IS SHOWN IN **BOLD**

Method	Params	FLOPS	Latency	IoU	mIoU
OccFormer [74]	132.5 M	825.8 G	0.348 s	36.50	13.46
Symphonies [63]	**59.31 M**	**611.9 G**	**0.314 s**	41.44	13.44
Ours	73.71 M	646.8G	0.324 s	**40.52**	**13.78**

architectures on our method. The evaluated backbones include ResNet-101-DCN [81], initialized from the FCOS3D [82] checkpoint; ResNet-50 [36], initialized from the MaskDINO [70] checkpoint; SwinL-DINO, with SwinL [68] as the backbone and initialized from the DINO [83] checkpoint; and SwinL-MD, with SwinL [68] as the backbone and initialized from the MaskDINO [70] checkpoint. As shown in Table VIII, using SwinL as the backbone further improves our method, achieving gains of 1.31 in IoU and 0.13 in mIoU compared to ResNet-50. In future work, we aim to incorporate more efficient and lightweight backbones to improve both efficiency and overall performance.

5) Complexity Analysis: In this section, we compare the performance of our method with OccFormer and Symphonies in terms of the accuracy metrics IoU and mIoU, and analyze the parameters (Params), theoretical computational costs (FLOPs) and latency. According to the data in Table IX, our method demonstrates competitive IoU and mIoU while maintaining relatively acceptable levels of model parameters, computational costs and latency. Inference latency is measured with a batch size of 1 on an RTX 3090 GPU. We attribute the increase in model parameters primarily to the object-to-scene information aggregation process. With further optimization, our method has the potential for real-time processing of outdoor scenes.

V. CONCLUSION AND FUTURE WORK

In this work, we present a novel monocular semantic scene completion framework that leverages a single RGB image to predict per-voxel occupancy status and corresponding semantic labels. We introduce a plug-and-play semantic-aware guided module that integrates task-aware semantic priors into image features, aiming to weaken the interaction with geometrically relevant image features associated with negative query proposals, while simultaneously enhancing the interaction with positive query proposals and areas of interest. Additionally, we propose an interactive refinement transformer block to iteratively updates query proposals by integrating the object-scene aggregation flow and scene-object aggregation flow. As evidenced by our experiments, our approach sets a new standard on two challenging datasets, while effectively ensuring the object generation in 3D scenes. We also conduct a series of ablation studies to validate the effectiveness of our module designs.

Future Works: Despite achieving promising results, we see opportunities for future enhancements by considering the following aspects. 1) We define a set of learnable object queries and then iteratively optimize them using the IRT block. However, due to the lack of instance-level annotations, omissions of objects in scenes still occur. Moreover, considering the imbalance in the number of objects in different images, setting a fixed number of object queries is not optimal. Therefore, we focus on further enhancements in the dynamic setting of object queries. 2) The pre-trained depth estimator often yields unreliable values. Thus, mitigating or circumventing this issue remains a significant challenge that merits further investigation. Additionally, when projecting voxel queries onto a 2D plane, depth ambiguity issues result in entangled feature aggregation. Therefore, inspired by DFA3D [84], we propose to utilize depth-aware 3D deformable attention to further enhance our method.

REFERENCES

[1] Y. Liu, Y. Cong, G. Sun, and Z. Ding, "Lifelong visual-tactile spectral clustering for robotic object perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 818–829, Feb. 2023.

[2] H. Xiao, Y. Li, W. Kang, and Q. Wu, "Distinguishing and matching-aware unsupervised point cloud completion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5160–5173, Sep. 2023.

[3] Z. Lu, B. Cao, and Q. Hu, "LiDAR-camera continuous fusion in voxelized grid for semantic scene completion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 29, 2024, doi: 10.1109/TCSVT.2024.3435045.

[4] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.

[5] X. Wang, Y. A. Sekercioglu, T. Drummond, E. Natalizio, I. Fantoni, and V. Frémont, "Fast depth video compression for mobile RGB-D sensors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 673–686, Apr. 2016.

[6] F. Wang, D. Zhang, H. Zhang, J. Tang, and Q. Sun, "Semantic scene completion with cleaner self," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 867–877.

[7] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9297–9307.

[8] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

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

[10] 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. (NeurIPS)*, vol. 36, Jan. 2023, pp. 64318–64340.

[11] A. Vobecký et al., "POP-3D: Open-vocabulary 3D occupancy prediction from images," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 50545–50557.

[12] 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/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 21729–21740.

[13] C. Sima et al., "Scene as occupancy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 8406–8415.

[14] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11138–11147.

[15] Q. Zhou, J. Cao, H. Leng, Y. Yin, Y. Kun, and R. Zimmermann, "SOGDet: Semantic-occupancy guided multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jan. 2023, pp. 7668–7676.

[16] Q. Ma, X. Tan, Y. Qu, L. Ma, Z. Zhang, and Y. Xie, "COTR: Compact occupancy transformer for vision-based 3D occupancy prediction," 2023, *arXiv:2312.01919*.

[17] W. Ouyang, X. Song, B. Feng, and Z. Xu, "OctOcc: High-resolution 3D occupancy prediction with octree," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, Mar. 2024, pp. 4369–4377.

[18] 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. (CVPR)*, Jun. 2023, pp. 9223–9232.

[19] S. Silva, S. B. Wannigama, G. Jayatilaka, M. H. Khan, and R. Ragel, "Unified spatio-temporal tri-perspective view representation for 3D semantic occupancy prediction," 2024, *arXiv:2401.13785*.

[20] Z. Yan et al., "Tri-perspective view decomposition for geometry-aware depth completion," 2024, *arXiv:2403.15008*.

[21] J.-C. Li, J.-G. Lu, M. Wei, H.-Y. Kang, and Q.-H. Zhang, "TPV-IGKD: Image-guided knowledge distillation for 3D semantic segmentation with tri-plane-view," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 10405–10416, Aug. 2024, doi: 10.1109/TITS.2024.3361163.

[22] M. Pan et al., "RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision," 2023, *arXiv:2309.09502*.

[23] C. Zhang et al., "OccNeRF: Advancing 3D occupancy prediction in LiDAR-free environments," 2023, *arXiv:2312.09243*.

[24] A. Hayler, F. Wimbauer, D. Muhle, C. Rupprecht, and D. Cremers, "S4C: Self-supervised semantic scene completion with neural fields," 2023, *arXiv:2310.07522*.

[25] H. Shi et al., "Offboard occupancy refinement with hybrid propagation for autonomous driving," 2024, *arXiv:2403.08504*.

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

[27] A.-Q. Cao and R. de Charette, "SceneRF: Self-supervised monocular 3D scene reconstruction with radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9387–9398.

[28] M. Pan et al., "UniOcc: Unifying vision-centric 3D occupancy prediction with geometric and semantic rendering," 2023, *arXiv:2306.09117*.

[29] 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.

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

[31] H. Jiang et al., "Symphonize 3D semantic scene completion with contextual instance queries," 2023, *arXiv:2306.15670*.

[32] 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.

[33] 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, doi: 10.1109/TITS.2023.3344806.

[34] R. Miao et al., "OccDepth: A depth-aware method for 3D semantic scene completion," 2023, *arXiv:2302.13540*.

[35] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

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

[37] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[38] M. Li, S. Li, X. Zhang, and L. Zhang, "UniVS: Unified and universal video segmentation with prompts as queries," 2024, *arXiv:2402.18115*.

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

[40] P. Tang et al., "SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction," 2024, *arXiv:2404.09502*.

[41] S. Wang et al., "Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation," 2024, *arXiv:2404.11958*.

[42] 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. 1746–1754.

[43] Y.-X. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image," 2018, *arXiv:1806.05361*.

[44] X. Chen, Y. Xing, and G. Zeng, "Real-time semantic scene completion via feature aggregation and conditioned prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2830–2834.

[45] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3D semantic scene completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7800–7809.

[46] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "ForkNet: Multi-branch volumetric semantic completion from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8607–8616.

[47] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 733–749.

[48] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3075–3084.

[49] M. Garbade, Y.-T. Chen, J. Sawatzky, and J. Gall, "Two stream 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 416–425.

[50] J. Li et al., "RGBD based dimensional decomposition residual network for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7685–7694.

[51] 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. (CVPR)*, Jun. 2020, pp. 3348–3356.

[52] Y. Liu et al., "3D gated recurrent fusion for semantic scene completion," 2020, *arXiv:2002.07269*.

[53] Y. Cai, X. Chen, C. Zhang, K.-Y. Lin, X. Wang, and H. Li, "Semantic scene completion via integrating instances and scene in-the-loop," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 324–333.

[54] S. Li, C. Zou, Y. Li, X. Zhao, and Y. Gao, "Attention-based multi-modal fusion network for semantic scene completion," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jan. 2020, pp. 11402–11409.

[55] W. Zhang, G. Liu, and G. Tian, "HHA-based CNN image features for indoor loop closure detection," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 4634–4639.

[56] B. Huang, J. Li, J. Chen, G. Wang, J. Zhao, and T. Xu, "Anti-UAV410: A thermal infrared benchmark and customized scheme for tracking drones in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2852–2865, May 2024.

[57] R. Cheng, C. Agia, Y. Ren, X. Li, and B. Liu, "S3CNet: A sparse semantic scene completion network for LiDAR point clouds," in *Proc. Conf. Robot Learn.*, Nov. 2021, pp. 2148–2161.

[58] Z. Li, G. Li, T. H. Li, S. Liu, and W. Gao, "Semantic point cloud upsampling," *IEEE Trans. Multimedia*, vol. 25, pp. 3432–3442, 2022.

[59] Z. Li, S. Liu, and G. Li, "PointELM: Fast point cloud classification using deep random mapping based extreme learning machines," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.

[60] P. Li, Y. Shi, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Semi-supervised implicit scene completion from sparse LiDAR," 2021, *arXiv:2111.14798*.

[61] M. Zhong and G. Zeng, "Semantic point completion network for 3D semantic scene completion," in *Proc. Eur. Conf. Artif. Intell.*, Aug. 2020, pp. 2824–2831.

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

[63] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "PointOcc: Cylindrical tri-perspective view for point-based 3D semantic occupancy prediction," 2023, *arXiv:2308.16896*.

[64] A.-Q. Cao, A. Dai, and R. de Charette, "PaSCo: Urban 3D panoptic scene completion with uncertainty awareness," 2023, *arXiv:2312.02158*.

[65] R. Song, R. Xu, A. Festag, J. Ma, and A. Knoll, "FedBEVT: Federated learning bird's eye view perception transformer in road traffic systems," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 958–969, Jan. 2024.
[66] J. Liu, Z. Cao, J. Yang, X. Liu, Y. Yang, and Z. Qu, "Bird's-eye-view semantic segmentation with two-stream compact depth transformation and feature rectification," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 11, pp. 4546–4558, Nov. 2023.
[67] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[68] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[69] 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.
[70] F. Li et al., "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3041–3050.
[71] X. Huang et al., "Tag2Text: Guiding vision-language model via image tagging," 2023, *arXiv:2303.05657*.
[72] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.
[73] 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 *Proc. Conf. Robot Learn.*, vol. 164, 2022, pp. 180–191.
[74] R. Zhang et al., "MonoDETR: Depth-guided transformer for monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9121–9132.
[75] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[76] Y. Li et al., "SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving," 2023, *arXiv:2306.09001*.
[77] Z. Chen and B. Li, "Bridging the domain gap: Self-supervised 3D scene understanding with foundation models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, Jan. 2023, pp. 1–14.
[78] 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.
[79] 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.
[80] G. Wang et al., "OccGen: Generative multi-modal 3D occupancy prediction for autonomous driving," 2024, *arXiv:2404.15014*.
[81] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[82] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 913–922.
[83] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.
[84] 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.

Haihong Xiao received the M.S. degree from Nanjing Agricultural University, Nanjing, China, in 2021. He is currently pursuing the Ph.D. degree with South China University of Technology. His research interests include 3D vision, point cloud processing, and scene representation learning.

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.

Hao Liu received the B.E. degree from the University of Electronic Science and Technology of China in 2016, the M.E. degree from the National University of Defense Technology in 2018, and the Ph.D. degree from Sun Yat-sen University in 2023. He is currently a Research Fellow with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include 3D deep learning and NeRF, particularly in 3D object detection and multi-object tracking.

Yuqiong Li received the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2010. He is currently a Senior Researcher with the Key Laboratory for Mechanics in Fluid Solid Coupling Systems, Institute of Mechanics, Chinese Academy of Sciences. His research interests include vehicle-terra mechanics, in-situ mechanical survey of lunar soil, artificial intelligence, and machine learning.

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 include 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 served as the General/Program Co-Chair for the Shape Modeling International Conference in 2022, the Symposium on Solid and Physical Modeling in 2022 and 2023, the Geometric Modeling and Processing Conference in 2014 and 2021, and the Conference on Computational Visual Media in 2020. He is serving/has served on the editorial boards of IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, *Computer Graphics Forum*, and *Computational Visual Media*.