Bi-SSC: Geometric-Semantic Bidirectional Fusion for Camera-based 3D Semantic Scene Completion

Yujie Xue*, Ruihui Li*, Fan Wu†, Zhuo Tang, Kenli Li, Mingxing Duan†
College of Computer Science and Electronic Engineering, Hunan University
{xueyj, liruihui, wufan, ztang, lkl, duanmingxing}@hnu.edu.cn

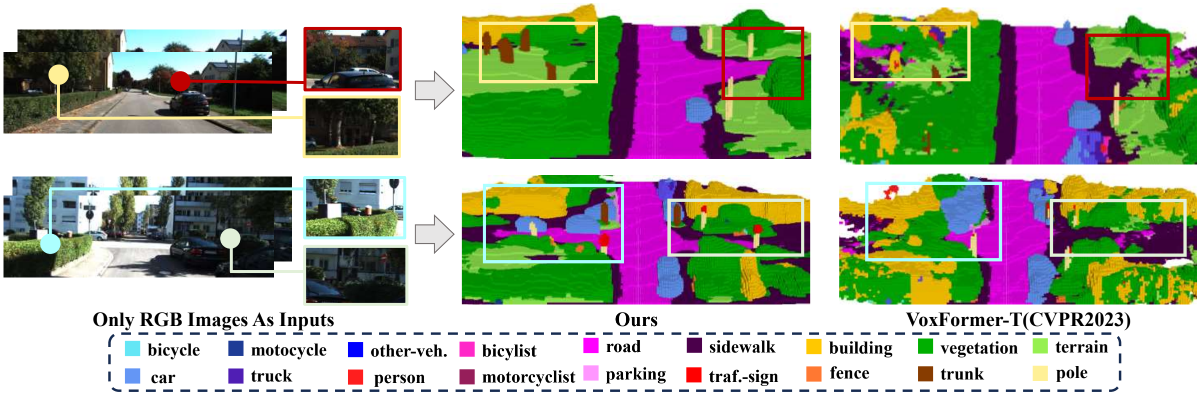

Figure 1. Given 2D image from the camera, our method is able to predict the complete 3D geometry of occluded objects and scenes. Clearly, our method excels not only in finely reconstructing the visible region, but also achieves better completion and segmentation of invisible and shaded areas, such as roads and poles in shaded areas, trees in shaded areas, and fine-grained profiles of cars overlaps. VoxFormer-T [25] also uses stereo depth performance for comparison, and the comparison benefits are marked with boxes.

Abstract

Camera-based Semantic Scene Completion (SSC) is to infer the full geometry of objects and scenes from only 2D images. The task is particularly challenging for those invisible areas, due to the inherent occlusions and lighting ambiguity. Existing works ignore the information missing or ambiguous in those shaded and occluded areas, resulting in distorted geometric prediction. To address this issue, we propose a novel method, Bi-SSC, bidirectional geometric semantic fusion for camera-based 3D semantic scene completion. The key insight is to use the neighboring structure of objects in the image and the spatial differences from different perspectives to compensate for the lack of information in occluded areas. Specifically, we introduce a spatial sensory fusion module with multiple association attention to improve semantic correlation in geometric distributions. This module works within single view and across stereo views to achieve global spatial consistency. Experimental results demonstrate that Bi-SSC outperforms state-of-the-art camera-based methods on SemanticKITTI, particularly excelling in those invisible and shaded areas.

1. Introduction

When faced with real-world objects of arbitrary shapes and infinite categories, the perception of the 3D environment is crucial for autonomous driving systems [11, 39]. It directly affects downstream tasks such as motion prediction and semantic map construction. However, constructing accurate and complete 3D information of the real world is notoriously difficult, since factors such as viewpoint occlusion or sparse noise.

To tackle these difficulties, 3D semantic scene completion (SSC) [41] is introduced, which formulates the problem as predicting the geometry and semantics of a scene through prior information. Subsequently, some excellent methods for realizing SSC with 3D information as input have been proposed [5, 6, 21, 46, 50, 61]. While LIDAR sensors offer relatively accurate depth measurement, cameras, despite being more cost-effective, can provide abundant visual information about the scene. The camera-based methods are emerging as an exciting alternative to LIDAR. This trend can be evidenced by MonoScene [3], which earliest approach that relied on monocular RGB images to infer 3D voxelized semantic scenes. However, it exposes the vulnerability of 2d-3d transformation, which is

[*] Equal contributed. [†] Corresponding authors.

inevitably subject to image occlusion and incomplete observation [15, 24, 26, 33]. Such as the reconstruction of roads obstructed by buildings is often unfeasible. In recent years,camera-based methods [16, 25, 51] seek to overcome this challenge through image modeling and dimensional transformation, but their inferences about the obscured area remain ambiguous. Additionally, although bird's eye view (BEV) perception provides a holistic representation of image features, and valuable support for obscured areas [27, 49]. Nevertheless, in recent works [20, 57], the holistic understanding of the 3D scene can hardly be recovered by using monocular or binocular BEV feature maps, especially for real-world obstacles with variable shapes.

Compared to existing methods, we consider building fine-grained 3D representations by integrating geometric and semantic features. This design is motivated by two factors: Firstly, images provide spatial distribution and semantic information of objects in the scene, the neighboring structures within the single view can be employed to predict the structure in the invisible region. Through geometric-semantic interaction, objects such as the pole covered by the car in Fig. 1 can be reconstructed. Furthermore, the spatial difference across different view images can alleviate the occlusion issue, thus, implementing cross-fusion of dual-view information can generate detailed and complete 3D scenes. Such as the precise outlines of overlapping vehicles in Fig. 1.

After revisiting the occlusion and illumination challenges present in SSC, we propose Bi-SSC in this paper, a framework that end-to-end bi-interactive feature framework. This approach utilizes two branches to preserve global information while incorporating geometric semantics to capture occlusion details. Building on the research of LSS [33], we specifically develop the *spatial sensory fusion*, which leverages multi-sensory integration and masks to query the neighboring structure and semantic information of visible objects. This improves the efficiency of transferring information from the image domain to the scene domain. Furthermore, we propose the *Cross-view Fusion* module to address the fusion bias of binocular features. By interacting with stereo matching, this module propagates and interacts with features from the left and right eyes, enhancing the global representation and enabling fine-grained semantic inference within specific occluded regions. In summary, the key contributions are as follows:

- We propose Spatial Sensory Fusion to lift image occlusion region into reliable geometric and semantic information, effectively compensate performance errors caused by occlusion.
- A method of Cross-view Fusion that propagates feature advantages to enhance scene representation.
- Experiments show that our Bi-SSC achieves state-of-the-art results of 16.73 mIoU and 45.10 IoU on the SemanticKITTI benchmark, outperforming all camera-only baseline methods.

2. Related Works

Semantic Scene Completion. Semantic Scene Completion restores the complete scene by understanding the objects and semantic relations, and predicting semantic information of the missing parts [41]. Previously, interpolation techniques for low-level features [8, 30] were used to extract information from the image for simple interpolation. However, these methods [9, 18, 29, 30, 42] are often inadequate when dealing with complex scenes, as they lack the ability to comprehend the semantic information. Therefore, recent work [37, 38, 45] has started to rely on deep learning to learn priors from large-scale datasets. Considering the 3D nature of the SSC task, researchers have directly employed 3D input data to enhance algorithm performance. Some of the studies [36, 37, 59] used point clouds to project features in the view space, and some approaches combine generative modeling [22, 34] to enhance the quality of the completion results [10, 48, 55], but their focus was solely on the scene completion task. Recently, some research on 3D SSC [37, 43], JS3C-Net [48] introduced a point-voxel interaction module to facilitate knowledge fusion between semantic segmentation and the scene completion task. SSA-SC [50] merged the semantic information from the segmentation branch into the completion branch. However, the usage of extensive 3D convolutions in the 3D methods make the model less efficient and cumbersome. In our work, we concentrate on extracting more precise geometric information from 2D data to aid the SSC task.

Camera-based 3D perception. The cost-effective and easy deployment of camera-based perception has attracted significant attention in SSC. The rich color information in images can extract comprehensive contextual details, assisting algorithms to accurately comprehend the scene [7, 23]. Several works have recently been proposed for SSC from RGB image [16, 17, 25, 27], such as VoxFormer [25], which utilizes deformable cross-attention to align occupancy positions with multi-frame image features and subsequently refines voxel features through deformable self-attention. TPVFormer [16] presents a transformer-based encoder that elevates image features to 3D TPV space. However, they are susceptible to the lack of fine-grained semantic information in voxels, resulting in inferior performance.

The BEV representation is adept at presenting the geometric configuration of the scene and the distribution of objects, enabling more effective utilization of visual information in building the scene. To transform image features into BEV features, researchers have employed the LSS [33] framework along with subsequent investigations [14, 35, 56, 60]. These studies involve projecting depth features from images at different perspectives onto 3D space. Other outstanding studies [1, 24, 58] focus on 3D object detection, such as BEVFormer [26], recommend a spatio-temporal converter that aggregates features from multiple image frames using variable attention mechanisms. For SSC, such as OccFormer [57] and StereoScene [20] use BEV perception to mitigate the effects of perspective transformations, such that enhance spatial understanding. Regrettably, the current BEV approaches with limited view challenges in understanding occluded areas for SSC.

Stereo matching based 3D perception. Driven by the relentless advancement of deep learning, stereo-matching methods have enhanced their effectiveness, thereby leading to substantial improvements in various 3D tasks [19, 40, 52]. Stereo matching methods can be broadly categorized into 2D CNN-based approaches [44, 47] and 3D CNN-based approaches [4, 31]. GwcNet [13] introduces grouped correlations to boost feature similarity measurement for more precise customer counting. Meanwhile, GA-Net [54] employs a novel CNN-based feature depth aggregation layer to enhance depth prediction accuracy and optimize finer structure and object edges. In the context of SSC tasks, existing stereo-matching methods suffer from issues such as non-textured regions and occlusion, which impede the accuracy of predicted depth.

3. Methodology

3.1. Preliminary

Problem setup. With the stereo images I_l^{rgb},I_r^{rgb} as input, the aim is to predict the geometry and semantics of a scene within a specific range in front. The predicted outputs are represented using a voxel grid $\mathbf{Y} \in \mathbb{R}^{H\times W\times Z}$, where H,W,Z denote the length, width, and height of the voxel grid respectively. As for each voxel, it is either empty denoted by c_0 or occupied by one of the semantic classes in $\mathbf{C} \in \{c_0, c_1, ..., c_N\}$. Here N denotes the total number of semantic classes. To summarize,our objective is to train a model $\mathbf{Y} = \Theta(I_l^{rgb}$,$I_r^{rgb})$ that can generate a 3D semantic prediction $\mathbf{Y}$ that closely approximates the ground truth $\bar{\mathbf{Y}}$.

Design rationale. Motivated by visual region information interaction and perspectives spatial difference fusion, we propose a framework for bi-interaction semantic geometric. Firstly, we learn neighbour structure features from the image semantic branch and the geometric branch, to obtain accurate features and alleviate the problems caused by occlusion. Next, leveraging the interaction between enhanced features across stereo views to achieve global spatial consistency, to improve voxel characterization.

Overall Architecture. We integrate semantically rich BEV features from 2D images to construct 3D voxel features. Fig. 2 shows the overall framework of Bi-SSC. Extract 2D features from RGB images, then use the designed *Spatial Sensory Fusion* (SSF) to generate semantic aware geometric features in *Semantic Geometric Fusion* (SGF) module. Subsequently, these features are refined and interactively utilized in the *Cross-view Fusion* (CVF) module to propagate information across all features. Finally, the resulting BEV features are up-sampled to voxel features for SSC. The specific process is outlined as follows:

- Utilizing the 2D U-Net architectures as the backbone from RGB images, and obtaining left and right image features $\mathbf{F}_l, \mathbf{F}_r \in \mathbb{R}^{C\times H\times W}$, respectively.
- The geometric features $F_G \in \mathbb{R}^{C\times H\times W}$ and semantic features $F_S \in \mathbb{R}^{C\times H\times W}$ through the corresponding network respectively. Then use SSF to fuse F_G, F_S as semantic-aware geometric features $\mathbf{F}_{sag} \in \mathbb{R}^{D_{sag}\times H\times W}$.
- The CVF module refines and interacts with the left and right features, resulting in a comprehensive set of dual-view features $\mathbf{F}_{Dual} \in \mathbb{R}^{D_d\times H\times W}$. Concurrently, the stereo matching method generate depth features $\mathbf{F}_s \in \mathbb{R}^{D_s\times H\times W}$ for querying F_{Dual}. Ultimately, BEV features are obtained by Mutual Interactive Aggregation (MIA) learn stereo features and refined features, so that they update their respective advantages of different depth features.
- These features are subsequently fed into the 3D UNet for semantic segmentation and scene completion.

The rest of this section details our innovations, the SGF module in Sec. 3.2, the CVF module in Sec. 3.3, and the training loss in Sec. 3.4.

3.2. Semantic Geometric Fusion Module

Mining occluding area information and refining the feature are crucial for addressing camera-based SSC tasks. The objective of SGF is to enhance the geometric and semantic associations to infer information about the occluded region. Thus, we designed *Spatial Sensory Fusion* (SSF) into our model inspired by Agentformer [53]. In the following paragraph, we will delve into the details of this approach.

As illustrated in Fig. 3, after obtaining the image features, we get the geometric features F_G and semantic features F_S through the neural network and then they are used as inputs to the SSF. Specifically, the geometric features are directly used as the query Q, and the semantic features are used as keys K and values V. To enhance the representation capabilities within the input sequences, we utilize two sets of projections $W_{in}, W_{out} \in \mathbb{R}^{l\times d}$, to generate spatial representations of inter-image and out-image semantics. Then multiplied with the key to get the projected keys $K_{in}, K_{out} \in \mathbb{R}^{l\times d}$. Here, the notation $l = H \times W$ represents a sequence of feature lengths for the image, while d corresponds to the dimensionality of the geometric feature. This projection operation facilitates a thorough exploration of the interrelationships among the input sequences, enabling a more comprehensive extraction of vital information. Following this, Q conducts separate queries into K_{in} and K_{out} to merge their dual outputs. To simulate the

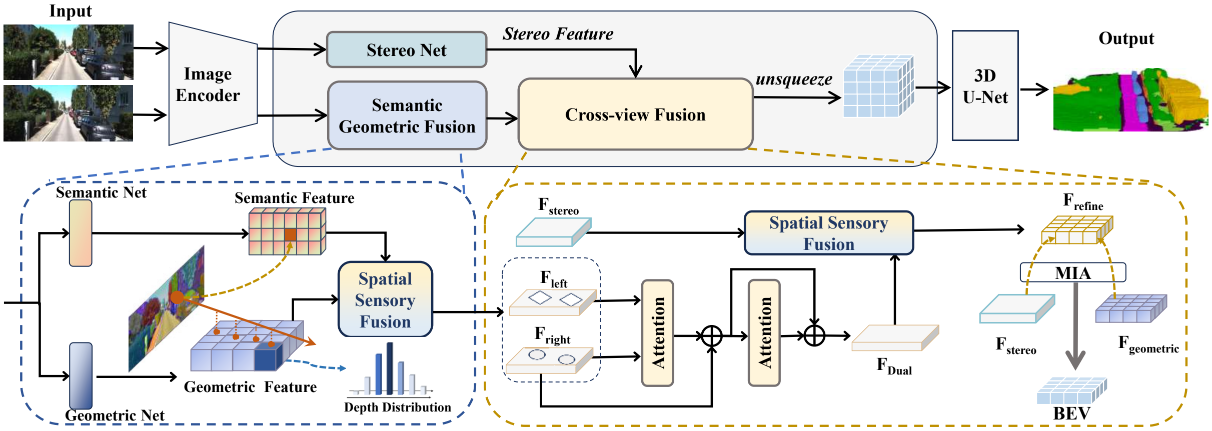

Figure 2. **Overall framework of Bi-SSC.** Given an input stereo image, the features extracted by 2D image encoder are respectively input to SGF and stereo network. In the SGF, our proposed SSF module is used to establish scene-level associations between geometric and semantic features. This is followed by CVF, where the fused features from both views are interacted to provide comprehensive global information. Afterward, the refined features and stereo features are sent into MIA to learn from each other. The resulting BEV features are upsampled into the output space, which enables accurate occupancy segmentation for each voxel.

impact of occluded regions and enhance the robustness of attention, we leverage an attention mask $M \in \mathbb{R}^{l\times d}$, as a feature intensifier in the feature space, which computes the consistency between each element A_{ij} in the attention weight matrix $A \in \mathbb{R}^{l\times d}$ and the sequence features. Here A_{ij} represents the attention weight between the query and two feature spaces in A. Through this operation, our model dynamically focuses on regions with similar features, resulting in greater precision in semantic segmentation. We express the process as:

$$M_{ij} = \mathbb{E}(\mathbf{N}_i = \mathbf{N}_j) \tag{1}$$

At this point, the output of the attention weight A_{ij} as

$$A = (QK_i^T) \odot M_{ij} + (QK_o^T) \odot (1 - M_{ij}) \tag{2}$$

where M_{ij} denotes each individual element of the selection mask, $\mathbb{E}$ represents the query function, and N corresponds to the number of receptors used for processing feature association. When belonging to the same attention receptors, M_{ij} is set to 1, otherwise M_{ij} is set to 0. That is, when the rows qk_i and columns qk_j of the mask M are the same, the attention weight A_{ij} is computed from the mask matrix. Notably, we use the same query to query different projection spaces and complement the missing positions based on the mask. This helps the model to pay attention to the feature expression at different locations in the input sequence.

Attention Score P aims to consider more important areas to improve output, and we incorporate it into the computation of the attention weight matrix. This approach explicitly leverages the significance of depth information in reconstructing the 3D scene and adeptly adjusts the feature weights within the attention mechanism. To achieve this, we convert each pixel value in the depth feature map into a probabilistic form using the *softmax* operation. Additionally, for each input image, the final attention score is determined by selecting the maximum value from each depth dimension. The formal representation of the process is:

$$p = max\left\{\frac{e^{P_i}}{\sum_{j=1}^{D} e^{p_j}}\right\} \tag{3}$$

where the $softmax$ is applied to the entire depth dimension D, Eq. (3) represents the calculation of the weight of each element. That is calculated by dividing the attention score of each element by the sum of all areas. This calculation effectively assigns relative importance to each element based on its corresponding attention score, enabling semantic awareness.

In the end, we update the previous attention matrix with the attention score as follows:

$$Attention(Q, K, V) = softmax(\frac{A}{\sqrt{d_k}} + \alpha P)V \tag{4}$$

where d_k represents the dimension of the query, and α serves as the balance coefficient. Thanks to the carefully crafted SSF design, the relationship between scene-level semantic geometry is effectively mined.

3.3. Cross-view Fusion Module

In our research, we introduce a novel Cross-view Fusion module, which aligns left and right fusion features according to spatial feature similarity. Specifically, we first map the left features F_l^{sag} into a query $Q_l \in \mathbb{R}^{C\times H\times W}$, the right features F_r^{sag} are mapped into keys $K_r \in \mathbb{R}^{C\times H\times W}$

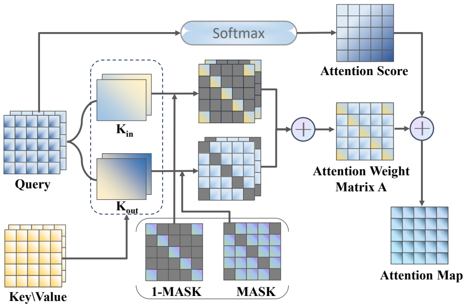

Figure 3. **A illustration diagram of the SSF.** Input the projection query, and the key/value undergoes separate projections before being queried. Then mask some features to simulate the effect of an obscured area, and the resulting attention weights are calculated independently. Finally, the attention scores generated from the query are added to obtain the attention map.

and values $V_r \in \mathbb{R}^{C \times H \times W}$, where C denotes the feature dimension. By leveraging attention, we compute the initial feature matching as A^l between features as follows:

$$A^l(Q_l, K_r, V_r) = softmax(\frac{Q_l K_r}{\sqrt{C}})V_r \tag{5}$$

The initial right feature is subsequently incorporated into the fusion as output F_{A^l}. Similarly, the initialized right feature F_r^{sag} mapping is transferred to a query $Q_r \in \mathbb{R}^{C \times H \times W}$, and the output F_{A^l} is mapped to the key $K_{A^l} \in \mathbb{R}^{C \times H \times W}$ and values $V_{A^l} \in \mathbb{R}^{C \times H \times W}$. At this point, the cross-attention of the right view with the features A^l can be computed as:

$$A^{Dual}(Q_r, K_{A^l}, V_{A^l}) = softmax(\frac{Q_r K_{A^l}}{\sqrt{C}})V_{A^l} \tag{6}$$

After several layers of attention, the dual-view features will be updated F_{Dual}. In order to alleviate the occlusion error of stereo matching, we combine the F_{Dual} and the stereo features F_s to get the refined features F_{refine} by using SSF:

$$F_{refine} = \mathit{SSF}(F_s, F_{Dual}) \tag{7}$$

Given the inherent bias present in left and right features, it is imperative to utilize these features in an efficient and stable manner for information exchange. Our CVF offers a mechanism to regulate the flow of information, progressively refining binocular image features.

After obtaining refined features F_{refine}, it will mutually enhanced with stereo features F_s and geometric features F_G. For superior aggregation, we utilize the wonderful Mutual Interactive Aggregation [20] module. Mathematically, the final BEV features F_{BEV} will be updated by the following general equation:

$$F_{BEV} = \mathit{MIA}(F_{refine}, F_s, F_G) \tag{8}$$

Specifically, the MIA selectively filters the most reliable information from the aggregated, stereo, and geometric features, following the standard protocol for StereoScene analysis. Note that we only show the formulation of the module for conciseness.

3.4. Loss Functions

In line with the learning objective of MonoScene [3] for semantic scene completion, we employ standard semantic loss ($\mathcal{L}_{sem}$) and geometric loss ($\mathcal{L}_{geo}$) to provide semantic and geometric supervision, respectively. Additionally, we incorporate class weighting loss ($\mathcal{L}_{ce}$) and binary cross-entropy loss ($\mathcal{L}_{depth}$) to promote a sparse depth distribution. To compute the final training loss, we simply sum these individual losses together as:

$$\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{geo} + \mathcal{L}_{ce} + \mathcal{L}_{depth} \tag{9}$$

4. Experiments

4.1. Experiments Setup

Dataset. We evaluated Bi-SSC on SemanticKITTI [2], the KITTI Odometry Benchmark [12] includes 22 complex, diverse and challenging outdoor driving scenarios. The voxels are generated through LIDAR scanning post-processing, where the ground truth semantic occupancy is represented as the 256×256×32 voxel grids, and each voxel size is 0.2m×0.2m×0.2m. The voxel grid is labeled with 21 classes (1 unknown, 1 free and 19 semantic). In the target output, SemanticKITTI generates ground truth semantic voxel grids by voxelizing a consistently registered semantic point cloud. SemanticKITTI can use the front camera and LIDAR points for SSC evaluation, but we use the binocular images obtained from cam2 and cam3 as inputs, since we are thinking about camera-only information. Moreover, to comprehensively evaluate the effectiveness of our model in complex scenarios, we introduce a SemanticKITTI-Complex dataset based on SemanticKITTI. This dataset was curated by five researchers who carefully handpicked 300 images from the SemanticKITTI validation set. These selected images specifically emphasize challenging conditions such as occluded areas and shaded areas, enabling a more rigorous assessment of our model's performance.

Evaluation metrics. For quantitative evaluations, we experimented with metrics that are widely used in the field of SSC. We utilize IoU (Intersection over Union) to evaluate the quality of scene completion and mIoU (mean Intersection over Union) to measure the performance of semantic segmentation, with higher values of both metrics implying better performance. Note that given the specific and challenging task of SSC, there is a strong interaction between IoU and mIoU, so the desired model should have excellent performance in both geometric completion and semantic segmentation.

Method	**Bi-SSC(Ours)**	**StereoScene** [20]	**VoxFormer-T** [25]	**OccFormer** [57]	**TPVFormer** [16]	**MonoScene** [3]
Input Modality	Stereo	Stereo	Stereo	Mono	Mono	Mono
IoU(%)	**45.10**	43.34	43.21	34.53	34.25	34.16
mIoU(%)	**16.73**	15.36	13.41	12.32	11.26	11.08
🟦 **car**(3.92%)	**25.00**	22.80	21.70	21.60	19.20	18.80
🟦 **bicycle**(0.03%)	1.80	**3.40**	1.90	1.50	1.00	0.50
🟦 **motocycle**(0.03%)	**2.90**	2.40	1.60	1.70	0.50	0.70
🟪 **truck**(0.16%)	**6.80**	2.80	3.60	1.20	3.70	3.30
🟦 **other-vehicle**(0.20%)	**6.80**	6.10	4.10	3.20	2.30	4.40
🟥 **person**(0.07%)	1.70	**2.90**	1.60	2.20	1.10	1.00
🟪 **bicylist**(0.07%)	**3.30**	2.20	1.10	1.10	2.40	1.40
🟪 **motorcyclist**(0.05%)	**1.00**	0.50	0.00	0.20	0.30	0.40
🟪 **road**(15.30%)	**63.40**	61.90	54.10	55.90	55.10	54.70
🟪 **parking**(1.12%)	**31.70**	30.70	25.10	31.50	27.40	24.80
🟪 **sidewalk**(11.13%)	**33.30**	31.20	26.90	30.30	27.20	27.10
🟥 **other-grnd**(0.56%)	**11.20**	10.70	7.30	6.50	6.50	5.70
🟨 **building**(14.10%)	**26.60**	24.20	23.50	15.70	14.80	14.40
🟧 **fence**(3.90%)	**19.40**	16.50	13.10	11.90	11.00	11.10
🟩 **vegetation**(39.3%)	**26.10**	23.80	24.40	16.80	13.90	14.90
🟫 **trunk**(0.51%)	**10.50**	8.40	8.10	3.90	2.60	2.40
🟩 **terrain**(9.17%)	**28.9**	27.00	24.20	21.30	20.40	19.50
🟨 **pole**(0.29%)	**9.30**	7.00	6.60	3.80	2.90	3.30
🟥 **traf.-sign**(0.08%)	**8.40**	7.20	5.70	3.70	1.50	2.10

Table 1. Semantic scene completion results on the SemanticKITTI [2] hidden test set with the state-of-the-art camera-based methods. We significantly outperform other methods in both IoU and mIoU, the best performing methods are marked in **bold**.

Implementation details. We crop RGB images to size 1280×384 and use image backbone network of EfficientNetB7 [3], set the input 3D feature volume size of the view transformer to 128×128×16, with 128 channels. The generated features are upsampled to 256×256×32 for full-scale evaluation. Unless otherwise specified, we have trained on the SemanticKITTI dataset with 30 epochs, using the AdamW [28] optimizer with an initial learning rate of 1e-4 and weight decay of 0.01. The learning rate is decayed by a multi-step scheduler. All models are implemented on PyTorch [32] using a Tesla A100 GPU.

Comparison Methods. We compare the best presently available models [3, 16, 20, 25, 57] that support 3D semantic scene completion. Among them are camera-based SSC methods for 2d-to-3d feature projection, such as MonoScene [3], VoxFormer [25], OccFormer [57], etc.

4.2. Main Results

Quantitative Comparison. We report the performance of Bi-SSC and RGB-inferred baselines for SemanticKITTI official benchmark (hidden test set), as shown in Tab. 1, the best results are shown in bold. Compared to State-of-the-art 2D methods, our method is greatly improved in terms of geometric completion and semantic segmentation. Significantly, our approach achieved superior performance over OccFormer, registering 4.41 mIoU (12.32 → 16.73, 35.8%) and 10.57 IoU (34.53 → 45.10, 30.61%) increase, respectively. VoxFormer-T Even using up to four temporal stereo image pairs as inputs, our mIoU and IoU still exceeded it, increasing by 3.32 mIoU (13.41 → 16.73, 24.76%) and 1.89 IoU (43.21 → 45.10, 4.37%), respectively. Such significant improvement is attributed to the fusion of geometric and semantic features in SSF to extract information from occluded regions, thereby alleviating the issue of visual blurring across the entire scene. For example, categories such as fence, building, and car have a lot of shielding in the scene, but they are still effectively improved. Furthermore, in comparison to StereoScene, Bi-SSC improves about +1.36 mIoU/+1.76 IoU on the SemanticKITTI dataset. Bi-SSC has the highest mIoU in almost all categories, it can be seen that almost all the classes get effective segmentation boosts. These results indicate that our attention module effectively captures the scene geometry without resorting to a simplistic increase in mIoU by decreasing the IoU values.

Qualitative results. In Fig. 6, we present the visualization of semantic scene completion prediction results on the SemanticKITTI validation set using Bi-SSC. To highlight the advantages of our method, we also include the results of OccFormer, VoxFormer-T, and their corresponding ground truth values (shown in the top row). Compared to the state-of-the-art VoxFormer-T[25], the spatial and semantic prediction outcomes of Bi-SSC exhibit significant improvements. This phenomenon is particularly noticeable in shielded areas and over extended distances. Such as in the first column of pictures, only our Bi-SSC is able to properly reconstruct the scene layout of the obscured road in the

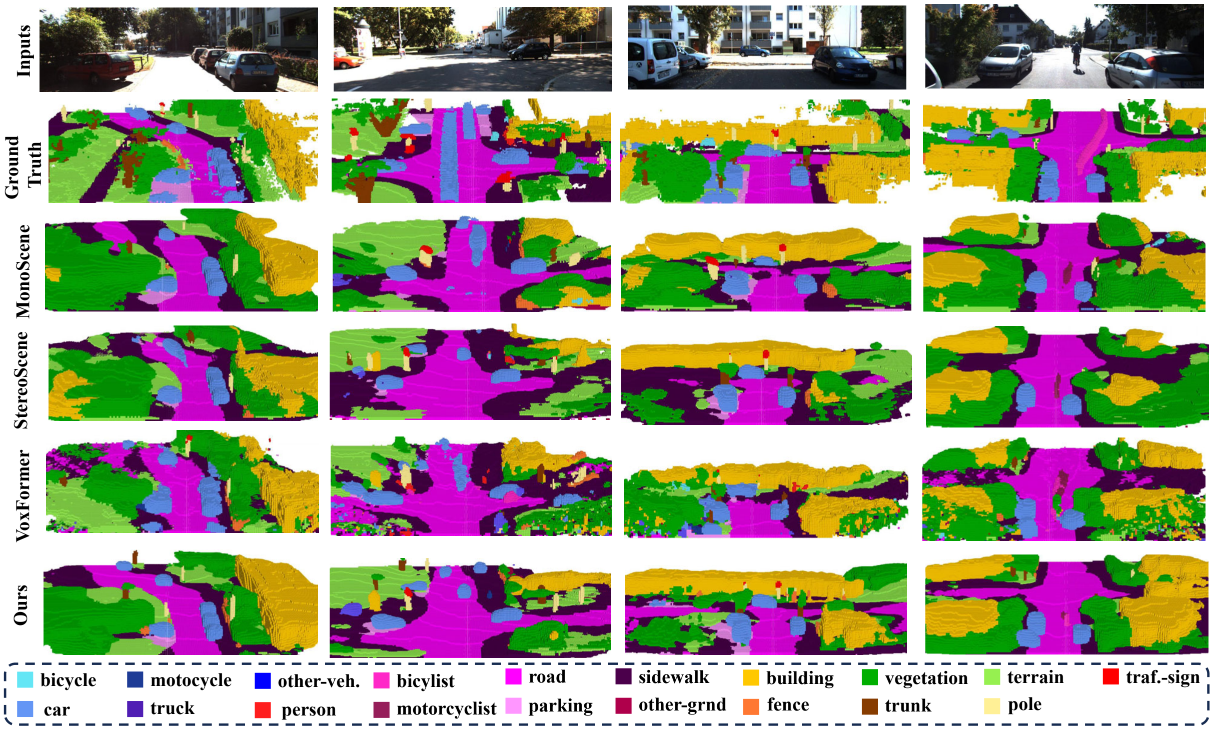

Figure 4. **Qualitative results from our method and others.** The input image perspectives are shown at the top, and then the 3D semantic occupancy results of Ground Truth, MonoScene [3], StereoScene [20], VoxFormer [25] and ours are shown in turn. Bi-SSC is able to better complement and segment the scene layout in large-scale autopilot scenarios. Also, Bi-SSC shows satisfactory results in the completion of small objects such as poles and occluded regions.

distance and the cars and trees in the shadows.

Our method outperforms MonoScene, OccFormer, and VoxFormer in comprehending scene-level layout and occluded regions. Moreover, Bi-SSC excels in recovering fine-grained structures and reasoning about interactions between neighboring semantic classes. For example, in the complex occlusion scene in the fourth column of pictures, our Bi-SSC can generate a more complete road extension and accurately segment the outline of each object. These advancements can be attributed to the effective aggregation of geometric with semantic features achieved by SSF.

4.3. Ablation Study

We conducted ablation experiments on the SemanticKITTI validation set to evaluate the impact of our Spatial Sensory Fusion, Cross-view Fusion module, and a contrast experiment for complex shaded areas.

Architectural Components. Tab. 2 presents a comprehensive analysis of how each architectural component contributes to achieving optimal results. The inclusion of the SSF module for feature fusion demonstrates a substantial enhancement in both geometric and semantic estimation, with a notable increase of 0.5 IoU and 1.04 mIoU, respectively. Furthermore, Thanks to the dual view receptive field and features aggregation, the CVF leads to significant improvements in geometric prediction (+0.87 IoU), while having a relative impact on semantic prediction (+0.54 mIoU). Finally, the incorporation of the MIA module further contributes to the overall accuracy, ultimately enhancing the performance in both geometric and semantic estimation.

Effectiveness of SSF module. We conducted in-depth ablation study on the SSF module to validate our design choices. The corresponding results are presented in Tab. 3. Initially, we removed the attention score design and compared it against other baseline methods. The analysis revealed that the removal of the attention score led to a noticeable impact on semantic segmentation, as indicated by the reduction in mIoU. This observation strongly supports the attention score is crucial for compensating geometric information at the scene level. Furthermore, Tab. 3 shows improved performance by varying query projection key values. It is worth noting can help improve the performance, due to the interaction of two different sets of distinct feature spaces. An interesting phenomenon is that only mIoU decreases when we ablate the Mask, which proves that it is useful for us to simulate the occluded area with the mask.

Architecture Components	IoU(%)	mIoU(%)
Ours	**44.88**	**16.39**
Ours w/o SSF	44.38	15.35
Ours w/o CVF	44.01	15.85
Ours w/o MIA	44.77	16.21

Table 2. **Ablation study for architecture.** Results are reported on SemanticKITTI val.

Method	IoU(%)	mIoU(%)
SSF	**44.88**	**16.39**
SSF w/o Attention Score	44.78	15.94
SSF w/o Dual Query	44.64	15.75
SSF w/o Mask	44.84	15.52
Attention	43.56	15.39

Table 3. **Ablation study for Spatial Sensory Fusion.** Our SSF module performs best, and each block has played its role.

Method	IoU(%)	decline(↓)	mIoU(%)	decline(↓)
MonoScene [3]	37.12	36.80 (0.32)↓	11.50	10.65 (0.85)↓
VoxFormer-T [25]	44.15	44.05 (**0.1**)↓	13.35	12.60 (0.75)↓
StereoScene [20]	43.85	42.54 (1.31)↓	15.43	13.93 (1.50)↓
Bi-SSC(Ours)	**44.88**	**44.78 (0.1)**↓	**16.39**	**16.0 (0.39)**↓

Table 4. **A comparison against the state-of-the-art method** in SemanticKITTI-Complex, where our approach exhibited no significant degradation in performance.

In comparison to the baseline attention mechanism, our SSF has been proven to be an effective attention mechanism for occluded regions in the SSC task.

Qualitative results of ablation studies. As illustrated in Fig. 5, compared with the full pipeline (a), the Voxformer-T (b) incorrect reconstruction of occluded roads. And removing SSF (c) will not learn neighboring structures (Car and road shelter structure in the yellow box), result in roads occluded by buildings cannot be reconstructed (blue circle). Removing CVF module distorts the geometry of the scene, and they both affect the details in the result. As in Fig. 5 (d), the road reconstruction lacks integrity. It proves the CVF module utilizes the spatial differences of different views (red box) to fill in road structures.

Our superiority over others in SemanticKITTI-Complex dataset. To ensure the validity of the experiment, other state-of-the-art methods were tested using their pre-trained models under identical conditions. The results in Tab. 4 demonstrate that our method surpasses other camera-based methods. Specifically, in this challenging dataset, Bi-SSC achieves a mIoU score of 16.0, which is 26.99% higher than VoxFormer-T and 14.86% higher than StereoScene, the most advanced methods in their respective categories. In addition, the decrease in mIoU for our method is only 0.39, compared to a decrease of 0.75 for VoxFormer-T and a significant decrease for StereoScene. This indicates that our method exhibits better robustness in challenging conditions. More importantly, Bi-SSC demonstrates the improvements in the area of occluded and shaded are significant, as shown in Fig. 6. Given the significance of accurate prediction in fuzzy environments, particularly in the field of autonomous driving, Bi-SSC should be more popular in this domain.

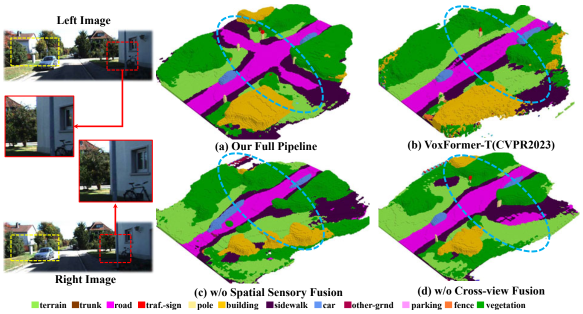

Figure 5. Visual results from the ablation study.

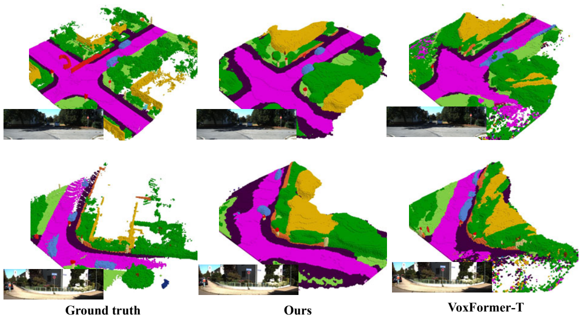

Figure 6. **Qualitative results in SemanticKITTI-Complex dataset.** Our approach better captures the layout of the scene, it reconstructs and estimates the geometry of the obscured roads and shaded areas of the car.

5. Conclusion

In this paper, we introduce Bi-SSC, an advanced camera-based framework for 3D semantic scene completion via geometric-semantic bidirectional fusion. We propose a Spatial Sensory Fusion that adeptly captures fine-grained features and scene-level information within two sets of image feature spaces. Moreover, we leverage Cross-view Fusion for dense geometric information fusion. As a result, Bi-SSC achieves a new SOTA performance in semantic scene completion on the SemanticKITTI, particularly excelling in those invisible and shaded areas.

Acknowledgement. The work is supported by the Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Major Project (2021ZD40300), National Natural Science Foundation of China (No.62202151) and the Fundamental Research Funds for the Central Universities.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 3

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 5, 6

[3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 5, 6, 7, 8

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 3

[5] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020. 1

[6] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021. 1

[7] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 2

[8] James Davis, Stephen R Marschner, Matt Garr, and Marc Levoy. Filling holes in complex surfaces using volumetric diffusion. In *Proceedings. First international symposium on 3d data processing visualization and transmission*, pages 428–441. IEEE, 2002. 2

[9] Raoul de Charette and Sotiris Manitsaris. 3d reconstruction of deformable revolving object under heavy hand interaction. *arXiv preprint arXiv:1908.01523*, 2019. 2

[10] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5431–5440, 2016. 2

[11] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, et al. Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends® in Robotics*, 8(1–2):1–224, 2020. 1

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5

[13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 3

[14] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021. 2

[15] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[16] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 2, 6

[17] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. *arXiv preprint arXiv:2306.15670*, 2023. 2

[18] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, page 0, 2006. 2

[19] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 3

[20] Bohan Li, Yasheng Sun, Xin Jin, Wenjun Zeng, Zheng Zhu, Xiaoefeng Wang, Yunpeng Zhang, James Okae, Hang Xiao, and Dalong Du. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 2, 3, 5, 6, 7, 8

[21] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 1

[22] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. Sp-gan: Sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2

[23] Shichao Li and Kwang-Ting Cheng. Joint stereo 3d object detection and implicit surface reconstruction. *arXiv preprint arXiv:2111.12924*, 2021. 2

[24] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2, 3

[25] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anand-

kumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1, 2, 6, 7, 8

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2, 3

[27] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 381–389, 2006. 2

[30] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 2

[31] Haesol Park and Kyoung Mu Lee. Look wider to match image patches with convolutional neural networks. *IEEE Signal Processing Letters*, 24(12):1788–1792, 2016. 3

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[33] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[34] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 366–383. Springer, 2020. 2

[35] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2

[36] Christoph B Rist, David Schmidt, Markus Enzweiler, and Dariu M Gavrila. Scssnet: Learning spatially-conditioned scene segmentation on lidar point clouds. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1086–1093. IEEE, 2020. 2

[37] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021. 2

[38] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 2

[39] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8):1978–2005, 2022. 1

[40] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4641–4650, 2017. 3

[41] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 1, 2

[42] Minhyuk Sung, Vladimir G Kim, Roland Angst, and Leonidas Guibas. Data-driven structural priors for shape completion. *ACM Transactions on Graphics (TOG)*, 34(6): 1–11, 2015. 2

[43] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2352–2360, 2022. 2

[44] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 3

[45] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 801–810. IEEE, 2020. 2

[46] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17651, 2023. 1

[47] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3

[48] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 2

[49] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. Bevheight: A robust framework for vision-based roadside 3d object detection. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21611–21620, 2023. 2

[50] Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3555–3562. IEEE, 2021. 1, 2

[51] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9455–9465, 2023. 2

[52] Xiaoqing Ye, Jiamao Li, Han Wang, Hexiao Huang, and Xiaolin Zhang. Efficient stereo matching leveraging deep local and context information. *IEEE Access*, 5:18745–18755, 2017. 3

[53] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 3

[54] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 3

[55] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 733–749, 2018. 2

[56] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2

[57] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 2, 3, 6

[58] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Jing Zhang, and Dacheng Tao. Bevsimdet: Simulated multi-modal distillation in bird's-eye view for multi-view 3d object detection. *arXiv preprint arXiv:2303.16818*, 2023. 3

[59] Min Zhong and Gang Zeng. Semantic point completion network for 3d semantic scene completion. In *ECAI 2020*, pages 2824–2831. IOS Press, 2020. 2

[60] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022. 2

[61] Hao Zou, Xuemeng Yang, Tianxin Huang, Chujuan Zhang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Up-to-down network: Fusing multi-scale context for 3d semantic scene completion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 16–23. IEEE, 2021. 1