Bridging Stereo Geometry and BEV Representation with Reliable Mutual Interaction for Semantic Scene Completion

Bohan Li[1,2] , **Yasheng Sun**[3] , **Zhujin Liang**[4] , **Dalong Du**[4] ,
Zhuanghui Zhang[4] , **Xiaofeng Wang**[5] , **Yunnan Wang**[1,2] , **Xin Jin**[2,*] and **Wenjun Zeng**[2]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]Tokyo Institute of Technology, Tokyo, Japan
[4]PhiGent Robotics, Beijing, China
[5]Chinese Academy of Sciences, Beijing, China
{bohan_li, wangyunnan}@sjtu.edu.cn, sun.y.aj@m.titech.ac.jp,
{zhujin.liang, dalong.du, zhuanghui.zhang}@phigent.ai,
wangxiaofeng2020@ia.ac.cn, {jinxin, wenjunzengvp}@eias.ac.cn

Abstract

3D semantic scene completion (SSC) is an ill-posed perception task that requires inferring a dense 3D scene from limited observations. Previous camera-based methods struggle to predict accurate semantic scenes due to inherent geometric ambiguity and incomplete observations. In this paper, we resort to *stereo matching* technique and *bird's-eye-view (BEV)* representation learning to address such issues in SSC. Complementary to each other, stereo matching mitigates geometric ambiguity with epipolar constraint while BEV representation enhances the hallucination ability for invisible regions with global semantic context. However, due to the inherent representation gap between stereo geometry and BEV features, it is non-trivial to bridge them for *dense prediction task* of SSC. Therefore, we further develop a unified occupancy-based framework dubbed **BRGScene**, which effectively **br**id**g**es these two representations with dense 3D volumes for reliable semantic **scene** completion. Specifically, we design a novel Mutual Interactive Ensemble (MIE) block for pixel-level reliable aggregation of stereo geometry and BEV features. Within the MIE block, a Bi-directional Reliable Interaction (BRI) module, enhanced with confidence re-weighting, is employed to encourage fine-grained interaction through mutual guidance. Besides, a Dual Volume Ensemble (DVE) module is introduced to facilitate complementary aggregation through channel-wise recalibration and multi-group voting. Our method outperforms all published camera-based methods on SemanticKITTI for semantic scene completion. Our code is available on https://github.com/Arlo0o/StereoScene.

1 Introduction

3D scene understanding is a fundamental task in computer vision [Roberts, 1963], facilitating a variety of applications such as autonomous driving, robotic navigation and augmented reality. Due to the limitations of real-world sensors such as restricted field of view, measurement noise, or sparse results, this task remains a challenging problem. To address this problem, 3D Semantic Scene Completion (SSC) [Roldao *et al.*, 2022] is introduced to jointly predict the geometry and semantic segmentation of a scene. Given its inherent 3D nature, most existing SSC solutions [Garbade *et al.*, 2019; Roldao *et al.*, 2020; Wu *et al.*, 2020] employ 3D geometric signals, in the form of occupancy grids, point clouds, or distance fields, as their model inputs. Although they provide insightful geometric cues, it requires costly sensors (e.g. LiDAR) alongside considerable manual labor entailed in their deployment. Hence, it is worth exploring an efficient and effective approach for high-fidelity SSC solely with cost-friendly portable cameras.

However, the absence of explicit 3D geometric information and incomplete observation pose large challenges to accurate geometry acquisition and reasonable hallucination in invisible regions [Cao and de Charette, 2022; Li *et al.*, 2023a]. Thus, previous camera-based SSC solutions [Cao and de Charette, 2022; Huang *et al.*, 2023] tend to utilize learning-based projection techniques to convert 2D image features into a 3D dense space, but their predictions inevitably fall short of capturing accurate geometry without explicit constraints. Later studies [Li *et al.*, 2023a] attempt to introduce depth information to augment query for reliable geometry prediction. But their results still struggle to hallucinate reasonable invisible regions without ensembling global semantic context.

As a pivotal technique in 3D vision applications, stereo matching leverages explicit epipolar constraint to establish pixel-level correspondence, which is suitable for reconstructing dense 3D scene geometry [Guo *et al.*, 2019]. On the other hand, the remarkable global robustness and hallucination ability of bird's-eye-view (BEV) representation, coupled

[*] Corresponding author.

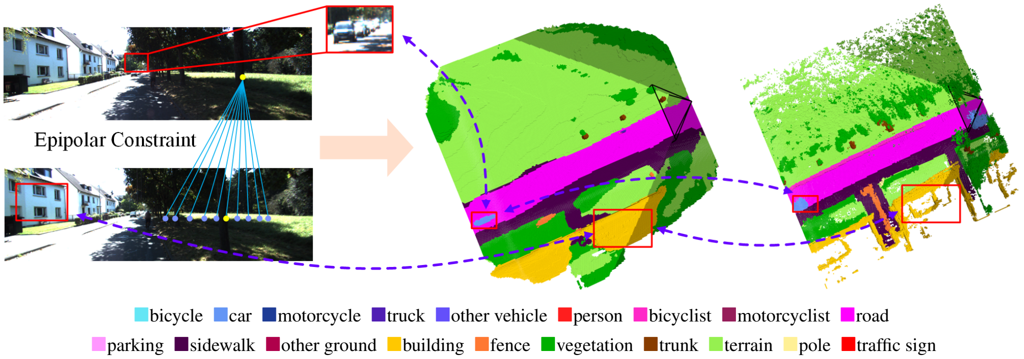

Figure 1: **Overview of the proposed BRGScene**. The figure illustrates the stereo inputs, SSC prediction results and ground truth from left to right. We can see that our method shows promising performance in completing semantic scenes, especially for those challenging distant small objects, as indicated by the car highlighted with a red box.

with rich context and global semantic information have led to its widespread utilization in the 3D object detection community [Philion and Fidler, 2020; Li *et al.*, 2022b; Roddick and Cipolla, 2020]. Inspired by the above two aspects, recent works have begun to simultaneously use stereo matching and BEV features for 3D perception. For instance, in the 3D object detection area, BEVStereo [Li *et al.*, 2022a] fuses monocular and temporal stereo depth maps before generating BEV features. BEVDepth [Li *et al.*, 2023b] improves geometry modeling with BVE representations by introducing extra depth supervision. These strategies significantly enhance the effectiveness of sparse perception tasks (e.g. 3D object detection) by focusing on coarse-grained predictions at the region level for common visual objects.

Despite remarkable advances in 3D object detection, it's non-trivial in the semantic scene completion (SSC) task to bridge the representation gap between stereo geometry and BEV features within a unified framework for pixel-level reliable prediction. This difficulty arises due to the structural variations among similar objects and the insufficient region-level coarse information for pixel-level semantics and geometry in complex real-world scenarios.

Given these concerns, we propose **BRGScene**, a framework that bridges stereo matching technique and BEV representation for fine-grained reliable SSC, and the results are shown in Figure 1. Our framework aims to fully exploit the potential of vision inputs with explicit geometric constraint of stereo matching [Guo *et al.*, 2019; Li *et al.*, 2021] and global semantic context of BEV representation [Philion and Fidler, 2020; Li *et al.*, 2023b]. Different from previous methods that focus on 2D features or depth maps [Li *et al.*, 2023a; Li *et al.*, 2023b], we propose to employ the dense 3D volumes of stereo and BEV representations for SSC.

Given the distinct nature of the two volumes, we devise a *Mutual Interactive Ensemble (MIE)* block to bridge the gap for fine-grained reliable perception. Specifically, a *Bi-directional Reliable Interaction (BRI)* module is designed to guide each volume to retrieve pixel-level reliable information. A confidence re-weighting strategy inspired by MVS [Chen *et al.*, 2020] is incorporated on top of the BRI module to further enhance the performance. Furthermore, a *Dual Volume Ensemble (DVE)* module is introduced to facilitate complementary aggregation with channel-wise recalibration and multi-group feature voting. Our contributions are summarized as follows: **1)** We propose a novel framework that resorts to both stereo and BEV representations with dense 3D volumes for precise geometry modeling and hallucination ability enhancement in SSC. **2)** To bridge the representation gap for fine-grained reliable perception, a *Mutual Interactive Ensemble* block is designed to take advantage of the complementary merits of the volumes in the two representations. **3)** Our proposed BRGScene outperforms state-of-the-art VoxFormer-T with a 14.5% relative improvement on the SemanticKITTI leaderboard.

2 Related Works

2.1 Semantic Scene Completion

Semantic scene completion is a dense 3D perception task that jointly estimates semantic segmentation and scene completion [Behley *et al.*, 2019; Cai *et al.*, 2021]. To provide additional texture or geometry information, some works [Cai *et al.*, 2021; Li *et al.*, 2019] exploit multi-modal inputs, such as RGB images coupled with geometric cues. Another slew of studies [Cao and de Charette, 2022; Li *et al.*, 2023a] aims to achieve semantic scene completion solely with camera-only inputs. For instance, MonoScene [Cao and de Charette, 2022] lifts a monocular image using 2D-3D projections and leverages 2D and 3D UNets for semantic scene completion. TPVFormer [Huang *et al.*, 2023] utilizes a tri-perspective view representation and attention mechanism for 3D scene understanding. VoxFormer [Li *et al.*, 2023a] employs a transformer-based framework where a sparse set of depth-based voxel queries are devised for scene structure reconstruction.

2.2 Stereo Matching Based 3D Perception

With the advances of deep convolution neural networks, the quality of depth predictions from stereo images [Poggi *et al.*, 2022] has steadily improved and led to a remarkable improvement in downstream 3D vision applications such as object detection, surface reconstruction and augmented reality. Recent methods like RAFTStereo [Deng, 2021] and CREStereo[Li *et al.*, 2021] utilize feature correlation to produce matching cost volume, which is subsequently optimized through sequential refinement modules for depth prediction. However, in conditions such as occlusion and large textureless regions, stereo matching prediction performance drops significantly.

2.3 Bird's-Eye-View Representation

The bird's-eye-view is a widely used representation in 3D object detection since it provides a clear depiction of the layout and strong hallucination ability from a top-down perspective [Philion and Fidler, 2020; Hu *et al.*, 2021]. Lift-Splat [Philion and Fidler, 2020] extracts BEV representations from an arbitrary number of cameras by implicitly unprojecting 2D visual inputs based on estimated depth distribution. BEVDepth [Li *et al.*, 2023b] leverages a camera-aware monocular depth estimation module to improve depth perception in BEV-based 3D detection. However, region-level coarse perception in the object detection task is not effective enough for the dense prediction task of SSC. In this work, our objective is to investigate reliable pixel-level prediction for high-quality semantic scene completion.

3 Methodology

We present our proposed BRGScene that aims to jointly infer dense 3D geometry and semantics solely from RGB images. In this section, we first introduce a hybrid occupancy-based SSC framework (Sec. 3.1), including problem formulation and architecture overview. Then we provide detailed construction of *dual volume* representations (Sec. 3.2). To bridge their representation gap for fine-grained reliable perception, we depict our devised ensemble block (Sec. 3.3). Finally, we introduce our SSC generator and training paradigm (Sec. 3.4).

3.1 The Proposed Framework of BRGScene

We propose BRGScene, a framework that bridges the stereo matching technique and the BEV representation for fine-grained reliable SSC.

Problem Formulation. Given a set of stereo RGB images $I^{rgb}_{Stereo} = \{I^{rgb}_l, I^{rgb}_r\}$, our goal is to jointly infer geometry and semantics of a 3D scene. The scene is represented as a voxel grid $\mathbf{Y} \in \mathbb{R}^{H \times W \times Z}$, where H, W, Z denote the height, width and depth in 3D space. Regarding each voxel, it will be assigned to a unique semantic label belonging to $C \in \{c_0, c_1, \cdots, c_M\}$, which either occupies empty space c_0 or falls on a specific semantic class $\{c_1, c_2, \cdots, c_M\}$. Here M denotes the total number of semantic classes. We would like to learn a transformation $\hat{\mathbf{Y}} = \Theta(I^{rgb}_{Stereo})$ to approach ground truth 3D semantics $\mathbf{Y}$.

Architecture Overview. The overall architecture of our proposed framework is illustrated in Figure 2. We follow a common paradigm [Cao and de Charette, 2022] that employs successive 2D and 3D UNets as backbones. The input stereo images I^{rgb}_{Stereo} are separately encoded by a 2D UNet into paired context-aware features $\mathbf{F}_l$ and $\mathbf{F}_r$. Then we leverage a *Stereo Constructor* to convert these features into a dense 3D volume $\mathbf{V}_{Stereo} \in \mathbb{R}^{D \times H \times W}$. In parallel, a *BEV Constructor* lift 2D features $\mathbf{F}_l$ of the left image to a latent BEV volume $\mathbf{V}_{BEV} \in \mathbb{R}^{D \times H \times W}$ alongside its context feature $\mathbf{C}_{BEV} \in \mathbb{R}^{C_b \times H \times W}$ following standard protocol of [Philion and Fidler, 2020]. Sequentially, the two-stream built volumes are bridged and aggregated to a new volume $\mathbf{V}_{ens}$ by a *Mutual Interactive Ensemble* block. Finally, the context features $\mathbf{C}_{BEV}$ splat along volume $\mathbf{V}_{ens}$ by outer-product, which will be fed to a 3D UNet for semantic segmentation and completion.

3.2 Dual Volume Construction

Unlike previous studies [Li *et al.*, 2023a; Li *et al.*, 2023b] which focus on 2D representations, we employ *3D volumetric representation* to resolve dense scene understanding. Specially, we introduce hybrid volumetric representations with stereo and BEV volumes to take full advantage of camera inputs.

2D Feature Extraction Backbone. For image feature extraction, 2D UNet with pre-trained EfficientNetB7 [Tan and Le, 2019] is leveraged to separately process left and right input images. Note that we utilize shared weights to encourage efficient correspondence feature learning.

Stereo Geometric Volume Constructor. With the obtained unary features $\mathbf{F}_l$ and $\mathbf{F}_r$ from the left and right images, *Stereo Constructor* targets to build a stereo depth volume $\mathbf{V}_{Stereo}$ by matching them with epipolar constraint. Specifically, group-wise correlation [Guo *et al.*, 2019] is first adopted to generate disparity cost volume. Formally,

$$D_{gwc}(d, x, y, g) = \frac{1}{N_c/N_g} \left\langle f^l_g(x, y), f^r_g(x - d, y) \right\rangle, \quad (1)$$

where $\langle \cdot, \cdot \rangle$represents the inner product, N_c is the channels of input features, N_g is the number of groups, f^l_g and f^r_g represent g^{th} left and right feature group, respectively. Afterward, the disparity volume is converted into a depth volume following [You *et al.*, 2020], which is formulated as:

$$z_{(u,v)} = \frac{f_u \times b}{D_{(u,v)}}, x = \frac{(u - c_u) \times z}{f_u}, y = \frac{(v - c_v) \times z}{f_v}, \quad (2)$$

where f_u and f_v represent the horizontal and vertical focal length, (c_u, c_v) is the camera center. Next, we employ 3D CNNs following [Guo *et al.*, 2019] for dimension reduction and finally squeeze the channel dimension to construct the dense stereo depth volume $\mathbf{V}_{Stereo} \in \mathbb{R}^{D \times H \times W}$.

BEV Latent Volume Constructor. Although the stereo constructor provides accurate estimation in matched regions, it struggles in extreme conditions where severe occlusion or high reflection happens. Unlike the stereo-based approaches relying on strict geometric matching, BEV representations are obtained by lifting an image I^{rgb} to a shared bird's eye space through 3D prior. Following [Philion and Fidler, 2020], we feed visual features $\mathbf{F}_l$ to a neural network and obtain a latent depth distribution $\mathbf{V}_{BEV} \in \mathbb{R}^{D \times H \times W}$ with its associated context features $\mathbf{C}_{BEV} \in \mathbb{R}^{C_b \times H \times W}$. Since this distribution is essentially a voxel grid that stores the probability of all possible depths, we denote it as BEV latent volume.

3.3 Mutual Interactive Ensemble

To achieve fine-grained reliable perception, the Mutual Interactive Ensemble (MIE) block is introduced to bridge the

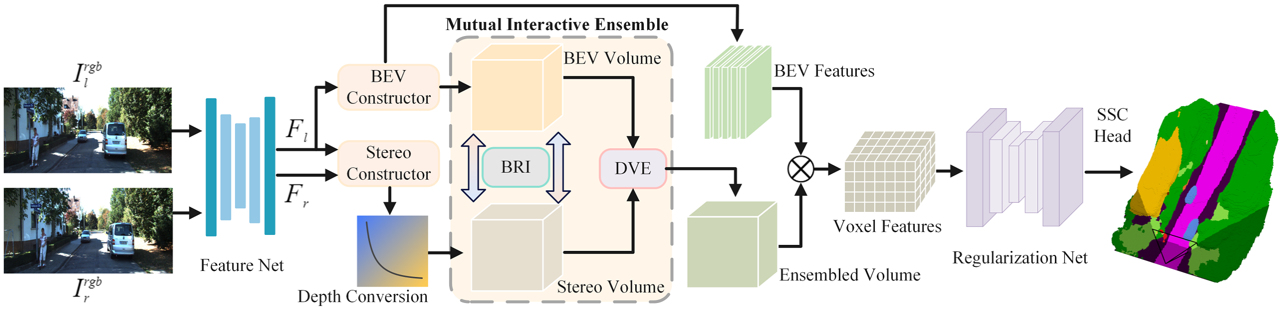

Figure 2: **Overall framework of our proposed BRGScene**. Given input stereo images, we employ 2D UNet to extract image features. The BEV latent volume and stereo geometric volume are constructed by a *BEV Constructor* and a *Stereo Constructor*, respectively. To bridge their representation gap for fine-grained reliable perception, a *Mutual Interactive Ensemble* block is proposed to take advantage of complementary merits of the volumes.

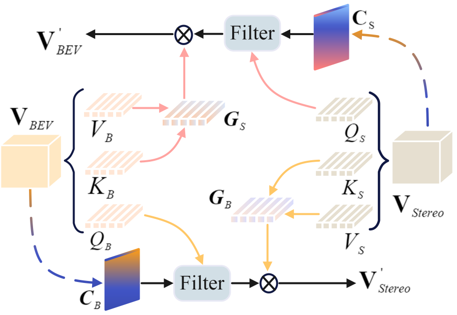

Figure 3: The structure of the proposed Bi-directional Reliable Interaction module, which is designed for pixel-level reliable geometry information interaction.

representation gap between stereo volume $\mathbf{V}_{Stereo}$ and BEV volume $\mathbf{V}_{BEV}$ by mutually reinforcing each other and integrating their respective potentials at pixel level.

Bi-directional Reliable Interaction

For pixel-level reliable interaction, we propose an initial interactive stage that selectively retrieves dependable information alongside its counterpart volume. More specifically, a Bi-directional Reliable Interaction (BRI) module as shown in Figure 3 is devised to interactively guide reliable predictions of its contrary side through a cross-attention mechanism. For stereo volume $\mathbf{V}_{Stereo}$, we first obtain its query Q_S, key K_S and value V_S by flattening in spatial and depth dimensions following standard protocol [Wang *et al.*, 2018; Liao *et al.*, 2022]. Similarly, the BEV volume $\mathbf{V}_{BEV}$ is forwarded and its query, key and value are denoted as Q_B, K_B, V_B, respectively.

Then we construct the interacted volume with the cross-attention operation. To reduce computational and memory consumption, we follow [Katharopoulos *et al.*, 2020; Kitaev *et al.*, 2020] to compute linear cross-attention. Specially, the interacted BEV volume $\mathbf{V}'_{BEV}$ is obtained by:

$$\begin{aligned}\mathbf{V}'_{BEV} &= CrossAtt(Q_S, K_B, V_B) \\ &= \phi_q(Q_S)\mathbf{G}_B = \phi_q(Q_S)(\phi_k(K_B)^T V_B),\end{aligned} \tag{3}$$

where ϕ_q and ϕ_k denote the softmax function along each row and column of the input matrix, respectively. $\mathbf{G}_B$ represents global contextual vectors of the BEV representation. In this way, $\mathbf{V}'_{BEV}$ retrieves the relevant aspects of the stereo sides, thereby providing an alternative perspective on the feature importance of the BEV side. Likewise, its opposite interacted volume $\mathbf{V}'_{Stereo}$ is computed by $CrossAtt(Q_B, K_S, V_S)$ to encourage reliable geometry information exchange.

Depth Confidence Filtering. In order to further retrieve pixel-level reliable information for dense prediction, we develop a depth confidence filtering strategy, which explicitly takes advantage of the involved reliable geometry information behind the volume. We aim to utilize its depth confidence information to enforce the cross-attention operation similar to [Chen *et al.*, 2020]. Particularly, to project the volume to a confidence map $\mathbf{C}_S$, we first adopt $softmax$ to convert depth cost value d_i into a probability form, and then take out the highest probability value among all depth hypothesis planes along the depth dimension as the prediction confidence. The process is formally written as:

$$\mathbf{C}_S = WTA(\phi(\mathbf{V}_{Stereo})) = WTA\left\{\frac{\exp(d_i)}{\sum_{j=1}^{D_{max}}\exp(d_j)}\right\}, \tag{4}$$

where the $softmax$ is applied across the depth dimension and WTA represents winner-takes-all operation. D_{max} denotes the length of the depth dimension.

Next, we revisit the cross-attention operation in Equation 3 and construct pixel-level reliable retrieval with the confidence map $\mathbf{C}_S$ to identify the criteria for an optimized formulation:

$$CrossAtt(Q_S, K_B, V_B) = \phi_q(Q_S) \odot \mathbf{C}_S(\phi_k(K_B)^T V_B), \tag{5}$$

where $\odot$ represents the element-wise product, through which the reliable geometry information is preserved while low-confidence information is suppressed.

Dual Volume Ensemble

With the interacted volume representations $\mathbf{V}'_{Stereo} \in \mathbb{R}^{1\times D\times H\times W}$ and $\mathbf{V}'_{BEV} \in \mathbb{R}^{1\times D\times H\times W}$, the primary objective of this module is to leverage their strengths and facilitate mutually beneficial complementation.

As illustrated in Figure 4, the DVE takes as input concatenated features $\mathbf{V}'_{cat} = [\mathbf{V}'_{Stereo}, \mathbf{V}'_{BEV}] \in \mathbb{R}^{2\times D\times H\times W}$ and outputs ensembled volume $\mathbf{V}_{ens}$. Especially, the input $\mathbf{V}'_{cat}$ is

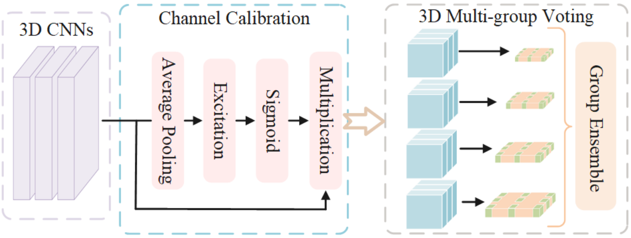

Figure 4: The structure of the proposed Dual Volume Ensemble module, which is devised for mutually beneficial aggregation.

first fed into residual 3D CNNs for regularization and channel adjustment, which generates transformed representation $\mathbf{V}_f \in \mathbb{R}^{C_f \times D \times H \times W}$. The transformed representation $\mathbf{V}_f$ is further processed by Channel-wise Recalibration and Multi-group Voting, which are described in detail below.

Channel-wise Recalibration. To fully exploit its contextual information [Hu *et al.*, 2018], we utilize average pooling to squeeze the information into a channel descriptor $\mathbf{z}_c$. More specifically, we shrink $\mathbf{V}_f$ along both the depth dimension D and spatial dimension $H \times W$:

$$\mathbf{z}_c = \frac{1}{D \times H \times W} \sum_{d=1}^{D} \sum_{i=1,j=1}^{H,W} \mathbf{V}_f(d,i,j), \tag{6}$$

Subsequently, an excitation block [Hu *et al.*, 2018] is leveraged to capture its channel-wise dependencies. Formally, the channel descriptor $\mathbf{z}_c$ is updated by two stacked bottleneck-shape convolutions with non-linear activation. Finally, the updated channel descriptor is employed to re-weight the previous transformed feature $\mathbf{V}_f$ along the channel dimension:

$$\mathbf{V}'_f = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z}_c)) \cdot \mathbf{V}_f. \tag{7}$$

where the W_1 and W_2 represent $1 \times 1 \times 1$ convolutions with dimensionality-reduction. The δ denotes standard GELU activation and the σ indicates the sigmoid gate.

Multi-group Voting. To further capture multi-scale context, we split $\mathbf{V}'_f$ into four groups and employ 3D atrous convolutions [Li *et al.*, 2019; Cao and de Charette, 2022] with different dilation rates. The contextual information in different groups exhibits distinct receptive fields, staying non-interfering with each other. Finally, we ensemble the multi-scale context between different voting groups to construct $\mathbf{V}_{ens}$:

$$\mathbf{V}_{ens} = \mathbb{P}\left\{\text{Concat}\left(A_g^i(\text{Split}_{channel}^{g1\sim g4}(\mathbf{V}'_f))\right)\right\}, \tag{8}$$

where A_g^i denotes 3D atrous convolution employed in the i^{th} context group. $\mathbb{P}$ is composed of point-wise convolution with GELU activation and group normalization. The point-wise convolution encourages channel-aware mixing by $1 \times 1 \times 1$ kernel size, through which the ensembled volume $\mathbf{V}_{ens}$ takes into consideration voting features of different aspects.

3.4 Semantic Scene Completion

To make use of this high-quality volume $\mathbf{V}_{ens}$ for semantic scene completion, we augment it with its associated context information $\mathbf{C}_{BEV}$. The extracted context information from input images is placed in a specific location of the bird-eye's representation by an outer product operation similar to [Philion and Fidler, 2020; Li *et al.*, 2023b]. Formally, the ensembled voxel features $\mathbf{F}_{vox} \in \mathbb{R}^{C_b \times D \times H \times W}$ is computed by:

$$\mathbf{F}_{vox} = \mathbf{C}_{BEV} \otimes \mathbf{V}_{ens}. \tag{9}$$

In this way, we are able to seamlessly blend the complementary benefits of stereo representation of precise geometry and BEV features of rich semantic context.

Semantic Segmentation Learning. Following [Cao and de Charette, 2022], we leverage 3D UNet to regularize the ensembled voxel features. Its output features are fed to a SSC head holding upsampling and a softmax layer for semantic occupancy prediction $\hat{\mathbf{Y}}$.

Network Training. We follow the basic learning objective of MonoScene [Cao and de Charette, 2022] for semantic scene completion. Standard semantic loss $\mathcal{L}_{\text{sem}}$ and geometry loss $\mathcal{L}_{\text{geo}}$ are leveraged for semantic and geometry supervision, while an extra class weighting loss $\mathcal{L}_{ce}$ is also added. To further enforce the ensembled volume, we adopt a binary cross entropy loss $\mathcal{L}_{depth}$ to encourage the sparse depth distribution. The overall learning objective of this framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{depth} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{sem}\mathcal{L}_{\text{sem}} + \lambda_{geo}\mathcal{L}_{\text{geo}}. \tag{10}$$

where several λs are balancing coefficients.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the proposed **BRGScene** on SemanticKITTI [Behley *et al.*, 2019] that is popularly used in previous studies. There are 22 driving outdoor scenes from the KITTI Odometry Benchmark [Geiger *et al.*, 2012], covering diverse and challenging autonomous driving situations. SemanticKITTI holds semantic annotations of LiDAR sweeps that are registered, aggregated and voxelized as 256×256×32 grid of 0.2m voxels. The target ground truth of each voxel grid is annotated as one of 21 classes (1 unknown, 1 free and 19 semantics). The SemanticKITTI benchmark provides both voxelized LiDAR scans and RGB images as model input options. We solely utilize RGB images since our main focus is to explore the portable camera-only signals as did in MonoScene [Cao and de Charette, 2022].

Implementation Details. We set the 3D UNet input to 128×128×16 (1:2) for efficient memory usage, whose feature will be upscaled to 1:1 at completion head by deconvolution operation. The λs are empirically set to 1. Our model is implemented on PyTorch. The model is trained for 30 epochs using the AdamW optimizer [Loshchilov and Hutter, 2017] with a learning rate of 1×10^{-4} and batch size set to 8.

Evaluation Metrics. Regarding quantitative evaluations, we conduct experiments on the typical metrics [Cao and de Charette, 2022] that have been widely employed in the field of SSC. Specially, we leverage **IoU** (Intersection over Union) to account for the scene completion (SC) task and **mIoU**

Methods	BRGScene(ours)	VoxFormer-T	VoxFormer-S	OccFormer	SurroundOcc	TPVFormer	MonoScene
Input	Stereo	Stereo-T	Stereo	Mono	Mono	Mono	Mono
IoU (%) ↑	**43.34**	43.21	42.95	34.53	34.72	34.25	34.16
mIoU (%) ↑	**15.36**	13.41	12.20	12.32	11.86	11.26	11.08
🟦 **car** (3.92%)	**22.80**	21.70	20.80	21.60	20.60	19.20	18.80
🟦 **bicycle** (0.03%)	**3.40**	1.90	1.00	1.50	1.60	1.00	0.50
🟦 **motorcycle** (0.03%)	**2.40**	1.60	0.70	1.70	1.20	0.50	0.70
🟪 **truck** (0.16%)	2.80	3.60	3.50	1.20	1.40	**3.70**	3.30
🟪 **other-veh.** (0.20%)	**6.10**	4.10	3.70	3.20	4.40	2.30	4.40
🟥 **person** (0.07%)	**2.90**	1.60	1.40	2.20	1.40	1.10	1.00
🟪 **bicyclist** (0.07%)	2.20	1.10	**2.60**	1.10	2.00	2.40	1.40
🟪 **motorcyclist** (0.05%)	**0.50**	0.00	0.20	0.20	0.10	0.30	0.40
🟪 **road** (15.30%)	**61.90**	54.10	53.90	55.90	56.90	55.10	54.70
🟪 **parking** (1.12%)	**30.70**	25.10	21.10	31.50	30.20	27.40	24.80
🟪 **sidewalk** (11.13%)	**31.20**	26.90	25.30	30.30	28.30	27.20	27.10
🟥 **other-grnd** (0.56%)	**10.70**	7.30	5.60	6.50	6.80	6.50	5.70
🟨 **building** (14.10%)	**24.20**	23.50	19.80	15.70	15.20	14.80	14.40
🟧 **fence** (3.90%)	**16.50**	13.10	11.10	11.90	11.30	11.00	11.10
🟩 **vegetation** (39.3%)	23.80	**24.40**	22.40	16.80	14.90	13.90	14.90
🟫 **trunk** (0.51%)	**8.40**	8.10	7.50	3.90	3.40	2.60	2.40
🟩 **terrain** (9.17%)	**27.00**	24.20	21.30	21.30	19.30	20.40	19.50
🟨 **pole** (0.29%)	**7.00**	6.60	5.10	3.80	3.90	2.90	3.30
🟥 **traf.-sign** (0.08%)	**7.20**	5.70	4.90	3.70	2.40	1.50	2.10

Table 1: **Quantitative results** on the SemanticKITTI test set. The top two performers are marked **bold**. Our method outperforms temporal stereo-based (Stereo-T) VoxFormer-T in terms of mIoU.

Methods	**Input**	**mIoU (%)** ↑	**Time (s)** ↓
SSCNet* (2017)	Stereo-PTS	10.31	**0.187**
LMSCNet* (2020)	Stereo-PTS	10.45	0.214
MonoScene* (2022)	Stereo	12.82	0.274
TPVFormer* (2023)	Stereo	13.06	0.313
OccFormer* (2023)	Stereo	13.57	0.338
VoxFormer-S (2023)	Stereo	12.35	0.256
VoxFormer-T (2023)	Stereo-T	13.35	0.307
BRGScene (ours)	Stereo	**15.43**	0.285

Table 2: **Evaluation results of stereo variants.** For MonoScene* and TPVFormer*, We employ stereo images as inputs. For SSCNet* and LMSCNet*, we leverage stereo depth net to generate pseudo point clouds (Stereo-PTS).

Methods	**Resolution**	**mAP** ↑	**NDS** ↑
BEVDet-Base (2021)	1600× 640	0.397	0.477
BEVDet4D-Base (2021)	1600× 640	0.426	0.552
PETR-R101 (2022)	1408× 512	0.357	0.421
BEVDepth-R101 (2023)	512× 1408	0.412	0.535
BRGScene (ours)	1600× 640	**0.451**	**0.563**

Table 3: **Quantitative results** of BEV Detection on the nuScenes validation set. We conduct preliminary experiments by employing the detection head.

(mean Intersection over Union) to measure the performance of the semantic scene completion (SSC) task, respectively. For both of these two metrics, higher values are desirable, where high IoU indicates accurate geometric prediction and high mIoU implies precise semantic segmentation.

4.2 Performance

Quantitative Comparison. Table 1 reports the performance of our BRGScene and other baselines on the SemanticKITTI test set. We compare our method with the best models [Li *et al.*, 2023a; Zhang *et al.*, 2023; Huang *et al.*, 2023; Wei *et al.*, 2023; Cao and de Charette, 2022] for semantic scene completion. BRGScene surpasses temporal stereo-based (Stereo-T) VoxFormer-T [Li *et al.*, 2023a] in terms of mIoU. Additionally, our approach outperforms stereo-based VoxFormer-S by a large margin in terms of geometric completion (42.95→43.34) and semantic segmentation (12.20→15.36). It's worth noting that our method demonstrates significant superiority in the prediction of small moving objects compared to VoxFormer-S, including bicycle (1.00→3.40), motorcycle (0.70→2.40), pole (5.10→7.00), etc. We ascribe such improvements to the ensemble of dual volume, which is critical for 3D geometry awareness. Besides, although VoxFormer-T employs up to 4 temporal stereo image pairs as inputs, our method has a significant advantage in mIoU, with a 14.5% relative improvement in terms of semantics.

Evaluation of Stereo Variants. To ensure fair comparisons, we also implement stereo variants of the baselines as shown in Table 2. For MonoScene* [Cao and de Charette, 2022] and TPVFormer* [Huang *et al.*, 2023], we employ left and right images to generate stereo-based predictions. Other LiDAR-based baselines including LMSCNet [Roldao *et al.*, 2020] and SSCNet [Song *et al.*, 2017] require 3D geometric inputs, so we adapt them with pseudo-3D inputs leveraging the same GwcNet [Guo *et al.*, 2019] used in our framework. Our proposed BRGScene efficiently outperforms all the other

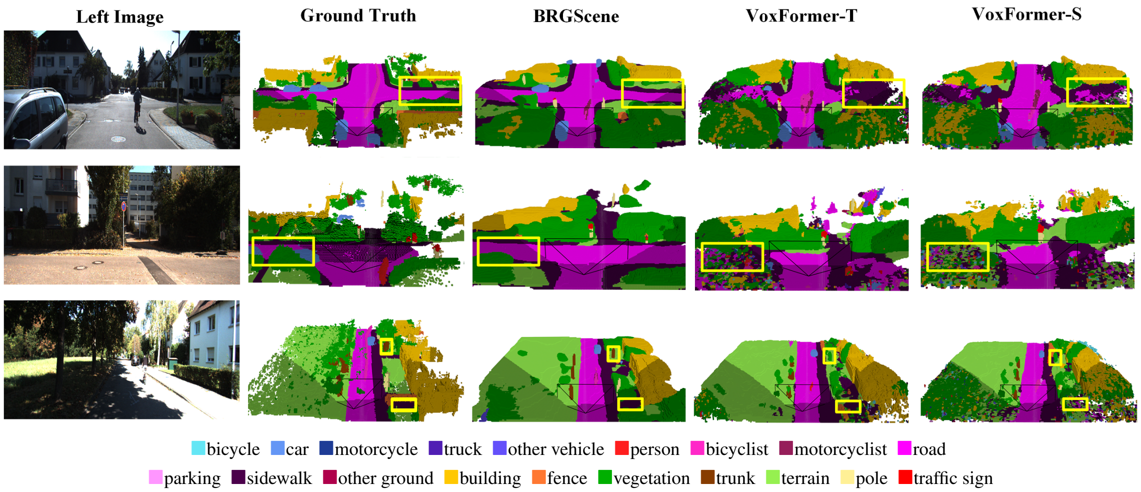

Figure 5: **Qualitative results** on the SemanticKITTI validation set. The overlay shadow areas at the bottom of semantic predictions denote unseen scenery out of the camera's field of view (FOV).

Dual Volume		MIE		IoU (%) ↑	mIoU (%) ↑
Stereo	BEV	BRI	DVE		
				33.87	9.92
✓				37.38	11.49
	✓			36.34	11.06
✓	✓			39.68	12.53
✓	✓	✓		43.04	14.64
✓	✓		✓	42.92	14.59
✓	✓	✓	✓	**43.85**	**15.43**

Table 4: **Ablation study** for architectural components.

methods. For more details on model complexity, please refer to the Supplementary Material.

Qualitative Comparison. As shown in Figure 5, we compare the visualization results of BRGScene and VoxFormer on the SemanticKITTI validation set. Due to the complexity of the real-world scenes and the sparsity of the labels, it is challenging to reconstruct the scenes accurately and completely. Compared to VoxFormer-T and VoxFormer-S, our method evidently captures better geometric representations for more complete and precise scene reconstruction (e.g. crossroads in rows 1,2) and generates more proper hallucinations outside FOV (e.g. shadow regions in rows 2,3).

BEV Detection Evaluation. We further conduct preliminary experimental results for BEV 3D detection on nuScenes validation set. Specifically, we adopt *BEVDet* [Huang *et al.*, 2021] as the baseline setting, and replace the *BEVDet* model with our proposed *BRGScene* while maintaining the detection head. Note that we adopt temporal inputs from current and previous images to construct the temporal volume, which replaces the original stereo volume. The preliminary results are in Table 3 , which show that our proposed method can also be applied to a more wide range of downstream tasks.

4.3 Ablation Study

We ablate our BRGScene on the SemanticKITTI validation set for semantic scene completion. The ablation study for the architectural components is shown in Table 4, which includes the Dual Volume and the Mutual Interactive Ensemble block. The baseline in the first row of the table is built by removing the Dual Volume and the MIE block.

Effect of the Dual Volume. For the ablation study on the dual volume, we build the framework with an individual volume to verify the effect. We find that after adding the stereo volume, the geometric perception ability of the framework improves obviously (+3.51 IoU), while the prediction performance of semantic scene segmentation is enhanced as well (+1.57 mIoU). The introduction of BEV volume also has an obvious impact on the semantic and the geometric aspects , boosting IoU by 2.47 and mIoU by 1.14, respectively.

Effect of the Mutual Interactive Ensemble. We further conduct architectural ablation to evaluate the impact of Mutual Interactive Ensemble (MIE) block as shown in Table 4. The BRI module can significantly improve the geometric and semantic estimations (+3.36 IoU, +2.11 mIoU) with efficient mutual interaction. Furthermore, we evaluate the DVE module by replacing the alternative naive concatenation, which leads to significant improvements in performance (+3.24 IoU, +2.06 mIoU). The aforementioned results validate that our complementary mutual interaction has a significant performance improvement compared to naive aggregation.

5 Conclusion

In this work, we propose BRGScene, a 3D Semantic Scene Completion framework that leverages both stereo and BEV representations to produce reliable 3D scene understanding results. To bridge the representation gap between the stereo volume and BEV volume for fine-grained 3D perception, a Mutual Interactive Ensemble block is proposed to incorporate complementary merits of the two dense volumes. Our BRGScene outperforms existing camera-based state-of-the-arts on the challenging SemanticKITTI dataset. We hope BRGScene could inspire further research in camera-based SSC and its applications in 3D scene understanding.

Acknowledgements

This paper is supported in part by NSFC under Grant 62302246 and ZJNSFC under Grant LQ23F010008.

References

[Behley *et al.*, 2019] 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 *ICCV*, 2019.

[Cai *et al.*, 2021] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, 2021.

[Cao and de Charette, 2022] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022.

[Chen *et al.*, 2020] Po-Heng Chen, Hsiao-Chien Yang, Kuan-Wen Chen, and Yong-Sheng Chen. Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Transactions on Image Processing*, 29, 2020.

[Deng, 2021] Lahav Lipson;Zachary Teed;Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. *3DV*, 2021.

[Garbade *et al.*, 2019] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *CVPRW*, 2019.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[Guo *et al.*, 2019] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[Hu *et al.*, 2021] 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 *ICCV*, 2021.

[Huang *et al.*, 2021] 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.

[Huang *et al.*, 2023] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023.

[Katharopoulos *et al.*, 2020] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020.

[Kitaev *et al.*, 2020] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020.

[Li *et al.*, 2019] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, 2019.

[Li *et al.*, 2021] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. *CVPR*, 2021.

[Li *et al.*, 2022a] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022.

[Li *et al.*, 2022b] 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 *ECCV*, 2022.

[Li *et al.*, 2023a] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, 2023.

[Li *et al.*, 2023b] 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 *AAAI*, 2023.

[Liao *et al.*, 2022] Jinli Liao, Yikang Ding, Yoli Shavit, Dihe Huang, Shihao Ren, Jia Guo, Wensen Feng, and Kai Zhang. Wt-mvsnet: window-based transformers for multi-view stereo. *NeurIPS*, 2022.

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

[Philion and Fidler, 2020] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.

[Poggi *et al.*, 2022] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Roberts, 1963] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[Roddick and Cipolla, 2020] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, 2020.

[Roldao *et al.*, 2020] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020.

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

[Song *et al.*, 2017] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[Wei *et al.*, 2023] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023.

[Wu *et al.*, 2020] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *3DV*, 2020.

[You *et al.*, 2020] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *ICLR*, 2020.

[Zhang *et al.*, 2023] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *ICCV*, 2023.