SPHERE: Semantic-PHysical Engaged REpresentation for 3D Semantic Scene Completion

Zhiwen Yang
Peking University
Wangxuan Institute of Computer Technology
Beijing, China
yangzhiwen@pku.edu.cn

Yuxin Peng*
Peking University
Wangxuan Institute of Computer Technology
Beijing, China
pengyuxin@pku.edu.cn

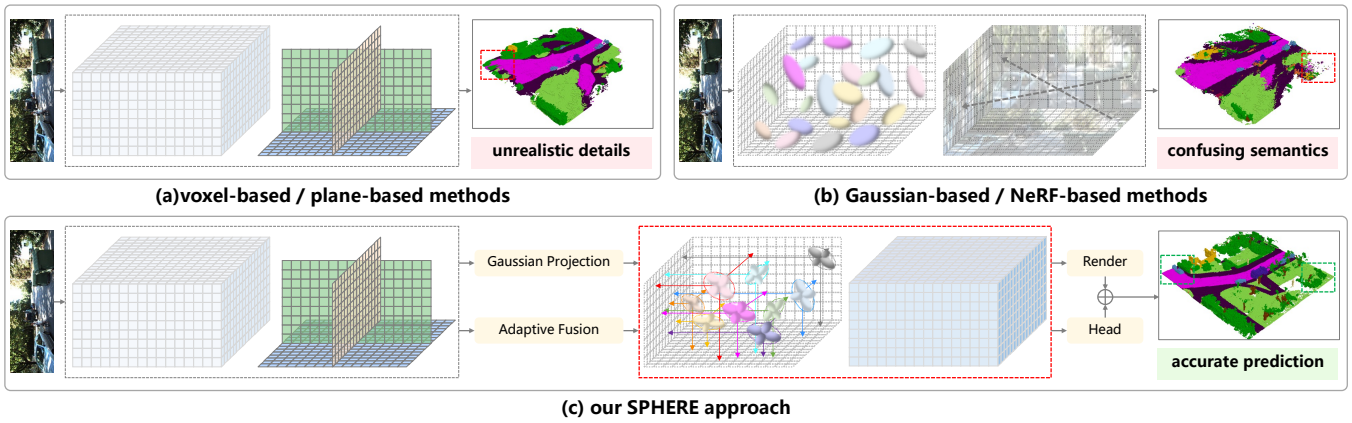

Figure 1: Comparison among different representations for SSC. (a) Voxel-based and plane-based methods demonstrate effectiveness in semantic accuracy, but struggle with modeling real-world physical regularities, leading to unrealistic details. (b) Due to spatial redundancy, Gaussian-based and NeRF-based methods suffer from high computation cost and slow convergence, resulting in confusing semantics. (c) Our SPHERE approach exploits semantic-geometry consistency via efficient integration of voxel and Gaussian representations, generating accurate SSC predictions.

Abstract

Camera-based 3D Semantic Scene Completion (SSC) is a critical task in autonomous driving systems, assessing voxel-level geometry and semantics for holistic scene perception. While existing voxel-based and plane-based SSC methods have achieved considerable progress, they struggle to capture physical regularities for realistic geometric details. On the other hand, neural reconstruction methods like NeRF and 3DGS demonstrate superior physical awareness, but suffer from high computational cost and slow convergence when handling large-scale, complex autonomous driving scenes, leading to inferior semantic accuracy. To address these issues, we propose the **S**emantic-**PH**ysical **E**ngaged **RE**presentation (**SPHERE**) for camera-based SSC, which integrates voxel and Gaussian representations for joint exploitation of semantic and physical information. First, the Semantic-guided Gaussian Initialization (SGI) module leverages dual-branch 3D scene representations to locate focal voxels as anchors to guide efficient Gaussian initialization. Then, the Physical-aware Harmonics Enhancement (PHE) module incorporates semantic spherical harmonics to model physical-aware contextual details and promote semantic-geometry consistency through focal distribution alignment, generating SSC results with realistic details. Extensive experiments and analyses on the popular SemanticKITTI and SSCBench-KITTI-360 benchmarks validate the effectiveness of SPHERE. The code is available at https://github.com/PKU-ICST-MIPL/SPHERE_ACMMM2025.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; **Reconstruction**.

Keywords

Semantic Scene Completion; Semantic-Geometry Consistency; Semantic Spherical Harmonics

ACM Reference Format:
Zhiwen Yang and Yuxin Peng. 2025. SPHERE: Semantic-PHysical Engaged REpresentation for 3D Semantic Scene Completion. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754882

*Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754882

1 Introduction

In autonomous driving systems, 3D Semantic Scene Completion (SSC) [33] is a crucial task for holistic scene perception, inferring voxel-level geometry and semantics for several downstream tasks. To deal with the inherent complexity of large-scale outdoor scenarios, LiDAR-based SSC methods [4, 14, 15, 31, 32, 40, 42, 45, 48] have been the predominant solutions, taking advantage of explicit 3D geometric structural information entailed in the point cloud data. Despite their effectiveness, the inherently high cost and limited scalability of LiDAR sensors constrain practical applications of LiDAR-based methods in real-world deployment. In light of this issue, multi-view images [27, 35, 43, 49] have garnered increasing attention as a more cost-effective and flexible input modality.

The pioneering camera-based SSC method, MonoScene [3], extracts 2D features from input RGB images and lifts them into dense 3D volume features through depth-based projection. Following this, existing camera-based SSC methods fall mainly into two paradigms: voxel-based and plane-based methods. Voxel-based SSC methods [19, 47] discretize 3D scenes into voxels and assign extracted image features to each voxel, and progress has been achieved in exploiting scene-from-instance feature interactions [12], and employing a dense-sparse-dense strategy for hybrid guidance [26]. In addition, plane-based SSC methods adopt Bird's-Eye-View (BEV) [8] and Tri-Perspective View (TPV) [10] features for efficient scene representations. Subsequent research further improves the quality of scene representation through camera-aware depth estimation and refinement [17], dynamic temporal stereo information [16], and explicit-implicit compact feature projection [25].

While existing SSC methods have demonstrated promising performance in voxel-wise semantic prediction, they frequently fail to capture the underlying physical regularities in the real world, leading to unrealistic geometric details in scene completion results, as illustrated in Figure 1 (a). On the other hand, neural reconstruction methods like NeRF [28] and 3DGS [13] excel at learning realistic geometry details, but suffer from high computational expense and slow convergence in large-scale complex autonomous driving scenes, resulting in confusing semantics, as shown in Figure 1 (b). The above challenges naturally raise a critical question: ***How can we achieve both accurate semantics and realistic geometry, efficiently and effectively?***

In this paper, we propose the *Semantic-PHysical Engaged REpresentation* (**SPHERE**) framework for 3D semantic scene completion as an alternative to answer this question, which combines voxel and Gaussian representations to jointly exploit semantic structures and physical regularities for accurate SSC prediction with accurate semantics and realistic geometry, as illustrated in Figure 1 (c). Our motivation derives from the observation that a large proportion of empty voxels in autonomous driving scenes cause severe spatial redundancy in neural representation methods, directly leading to high computational expense and slow convergence speed. In light of this, the *Semantic-guided Gaussian Initialization* (**SGI**) module is proposed to utilize dual-branch 3D scene representations to identify focal voxels with discriminative semantics as anchors, guiding efficient and effective Gaussian initialization. Then, to fully leverage the powerful physical structure modeling capability of Gaussian representations, we introduce the *Physical-aware Harmonics Enhancement* (**PHE**) module, which devises semantic spherical harmonics to model physical-aware contextual structures, then enhances semantic-geometry consistency via focal distribution alignment between voxel and Gaussian representations, thereby generating realistic semantic scene completion results. Extensive experiments and analyses on the SemanticKITTI and SSCBench-KITTI-360 datasets validate the effectiveness of SPHERE, demonstrating improved semantic accuracy and geometric fidelity over state-of-the-art methods.

In summary, the main contributions of our work are as follows:

- We propose the SPHERE framework, which efficiently integrates voxel and Gaussian representations to jointly exploit semantic information and physical regularities, generating SSC results with accurate semantics and realistic geometry.
- SGI leverages dual-branch 3D scene representations to identify focal voxels with discriminative semantics as anchors, enabling effective and efficient Gaussian initialization to alleviate spatial redundancy.
- PHE module incorporates semantic spherical harmonics for improved physical-aware contextual details, and promotes focal distribution alignment between voxel and Gaussian representations to enhance semantic-geometric consistency.

2 Related Work

In this section, we provide a brief review of pertinent literature related to our work, focusing on camera-based SSC methods and neural representation methods in 3D perception.

2.1 Camera-based Semantic Scene Completion

Voxel-based SSC methods discretize 3D space into voxels and assign image features to each voxel for scene perception. MonoScene [3] is a pioneering effort that reconstructs 3D scenes from RGB images by projecting image features to 3D space along optical rays. OccFormer [47] introduces a dual-path transformer architecture with a mask-wise prediction paradigm, effectively capturing spatial dependencies for improved completion accuracy. VoxFormer [19] presents a two-stage framework that integrates class-agnostic query proposals with class-specific semantic segmentation, facilitating the propagation of discriminative semantic information from seed voxels across the entire scene for enhanced completion. Symphonize [12] introduces a scene-from-instance framework that enhances the interplay between image and volumetric representations, enabling more precise scene completion. SGN [26] presents a one-stage SSC framework employing a dense-sparse-dense design, which enhances segmentation boundary clarity and improves the accuracy of 3D semantic scene completion. CGFormer [46] introduces a context- and geometry-aware voxel transformer that utilizes context-dependent queries and multiple 3D representations to enhance scene completion performance.

Plane-based SSC methods compress the 3D space along certain dimensions for efficient view transformation and scene representation. A pioneering approach in this domain is the Lift-Splat-Shoot (LSS) [30] framework, which establishes an end-to-end pipeline by first "lifting" images into feature frustums, then "splatting" these frustums onto a rasterized BEV grid, and finally "shooting" template

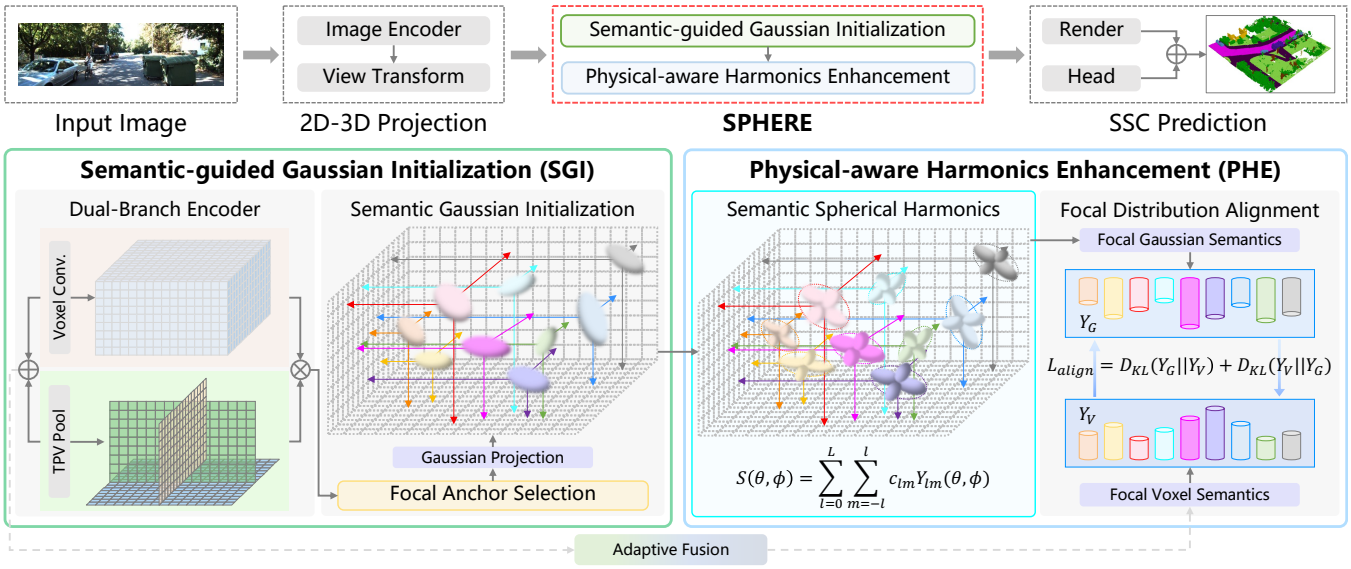

Figure 2: The overall architecture of our SPHERE approach. The Semantic-guided Gaussian Initialization (SGI) module leverages a dual-branch encoder to exploit local and global semantics, and then selects focal voxels as anchors for effective and efficient Gaussian initialization. The Physical-aware Harmonics Enhancement (PHE) module incorporates semantic spherical harmonics to model physical-aware contextual details, then enhances semantic-geometry consistency via focal distribution alignment between the voxel and Gaussian semantics.

trajectories into a BEV cost map for downstream tasks. Building upon the LSS paradigm, BEVDet [8] introduces a holistic BEV-based framework for scene understanding, comprising an Image-view Encoder, a View Transformer, a BEV Encoder, and a Task-specific Head. Subsequent achievements have been made in further enhancing the BEV-based representations, incorporating techniques such as inverse perspective mapping [7], angle and radius aware rasterization [22], and cross-attention layers with BEV queries [20]. BEVDepth [17] enhances depth learning by introducing a camera-aware depth estimation module alongside a depth refinement module, improving the accuracy of depth predictions. Building on this, BEVStereo [16] advances depth estimation by incorporating dynamic temporal stereo information. Beyond BEV, TPVFormer [10] introduces a Tri-Perspective View (TPV) representation, which decomposes voxel grids into three orthogonal planes, allowing efficient scene encoding.

Existing camera-based semantic scene completion methods have made great progress in predicting voxel-wise semantics, but fall short of the ability to capture physical regularities for realistic geometry details. In contrast, our SPHERE employs Gaussian representations around focal regions for the exploitation of physical-aware structures, facilitating realistic geometric details in the semantic scene completion results.

2.2 Neural Representation in 3D Perception

Neural representation techniques have emerged as powerful tools for modeling geometric details. Neural Radiance Field (NeRF) [28] represents the 3D scene with implicit continuous functions, mapping point coordinates and view directions to target properties, and employs ray-marching for volume rendering. 3D Gaussian Splatting (3DGS) [13] explicitly initializes a set of Gaussian distributions to represent the geometry and appearance within the scene, improving efficiency and scalability with a rasterization-based rendering pipeline. Omni-Scene [38] takes advantage of both pixel and volume-based representations and proposes an Omni-Gaussian representation for improved ego-centric reconstruction and novel view synthesis performance. RenderOcc [29] extracts a NeRF-style 3D scene representation and utilizes 2D rendering techniques for 3D supervision from 2D semantics and depth labels. GaussianFormer [9, 11] proposes image-to-Gaussian mapping with distribution-based initialization and Gaussian-to-voxel splatting with probabilistic Gaussian superposition for 3D semantic scene completion.

Although superior in learning physical structures and geometric details, neural representation methods suffer from high computational cost and slow convergence speed in autonomous driving scenes due to spatial redundancy. In contrast, our SPHERE approach identifies focal voxels as anchors to guide efficient and effective Gaussian initializations, alleviating spatial redundancy while enabling semantic scene completion with accurate semantics and realistic geometry.

3 Methodology

3.1 Overview

Figure 2 demonstrates the overall framework of our Semantic-Physical Engaged REpresentation (SPHERE) approach, consisting of two key components: (1) a Semantic-guided Gaussian Initialization (SGI) module that employs dual-branch 3D representation to select focal voxel anchors for efficient and effective Gaussian initialization;

(2) a Physical-aware Harmonics Enhancement (PHE) module that adopts semantic spherical harmonics for improved physical-aware contextual details and aligns focal voxel and Gaussian semantic distribution for enhanced semantic-geometry consistency.

Problem Setup. Given a pair of stereo images $I_l^{\text{rgb}}, I_r^{\text{rgb}}$, semantic scene completion (SSC) aims to infer the geometry and semantics of 3D scenes in front. The voxelized output is structured with grid size $Y \in \mathbb{R}^{X\times Y\times Z}$, where X, Y, Z correspond to the grid's length, width, and height, respectively. Each voxel is classified as either empty, denoted by c_0 or occupied by one of the semantic classes in $c \in \{c_1, \cdots, c_N\}$, where N represents the number of semantic classes. Essentially, SSC aims to train a model $V = \Theta(I_l^{\text{rgb}}, I_r^{\text{rgb}})$ that can generate a 3D semantic prediction V aligning closely with the ground truth $\bar{V}$.

3D Semantic Gaussian Representation. 3D semantic Gaussian representation initializes a set of P Gaussian distributions $\mathcal{G} = \{G_i\}_{i=1}^{P}$, where each G_i covers a local area determined by its mean m_i, scale s_i, rotation r_i, opacity α_i and semantic c_i. These Gaussian distributions contribute to voxel semantics in an additive manner:

$$y(\mathbf{x};\mathcal{G}) = \sum_{i=1}^{P} g_i(\mathbf{x}; m_i, s_i, r_i, \alpha_i, c_i) \tag{1}$$

where $y(\mathbf{x};\mathcal{G})$ denotes the predicted occupancy concerning all Gaussian distributions at location $\mathbf{x}$, and $g_i(x;\cdot)$ indicates the contribution of the i-th semantic Gaussian. The contribution is evaluated based on the Gaussian properties:

$$g_i(\mathbf{x}; G_i) = \alpha_i \cdot \exp(-\frac{1}{2}(x-m_i)^{\mathrm{T}}\Sigma_i^{-1}(x-m_i))c_i \tag{2}$$

$$\Sigma_i = RSS^{\mathrm{T}}R^{\mathrm{T}}, \quad S = \text{diag(s)}, \quad \text{R} = \text{q2r(r)} \tag{3}$$

where Σ_i is the covariance matrix, S_i represents the diagonal scale matrix, and R_i denotes the rotation matrix.

3.2 Semantic-guided Gaussian Initialization

Design rationale. The large proportion of empty voxels in autonomous driving scenes causes severe spatial redundancy in neural representation techniques like NeRF [28] and 3DGS [13] for modeling non-occupied areas, resulting in high computational expense and slow convergence speed. In light of this, the SGI module first leverages dual-branch 3D representations for global and local semantic information, and then identifies focal voxels as anchors for efficient and effective Gaussian initialization.

Dual-Branch Encoder. Notice that TPV [10] representations compress the full voxel representations along three axes respectively, thereby focusing more on global semantics, while dense voxel representations highlight local semantic details. We leverage both voxel and TPV representations for comprehensive aggregation of local and global semantic information. Given the initial 3D features $\mathbf{F}_{\text{3D}}$ projected from image features $\mathbf{F}_{\text{img}}$, the voxel representation is derived through multiple 3D convolution layers followed by a 3D pyramid network:

$$\mathbf{F}_{\text{voxel}} = \text{FPN}_{\text{3D}}\left(\text{Conv}_{\text{3D}}\left(\mathbf{F}_{\text{3D}}\right)\right) \tag{4}$$

where $\mathbf{F}_{\text{voxel}} \in \mathbb{R}^{C\times X\times Y\times Z}$ denotes the voxel representation with local semantic details. On the other hand, the TPV representations

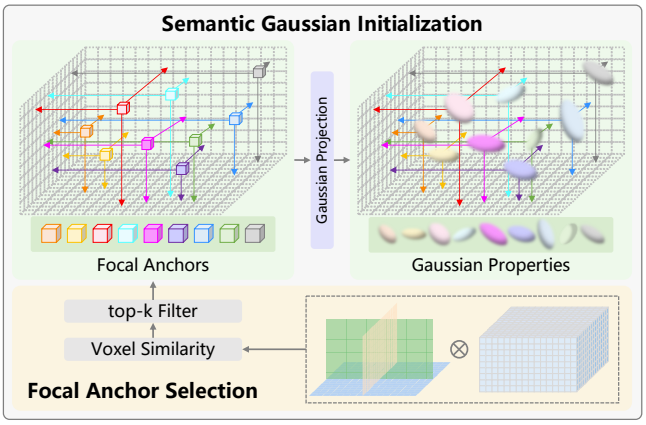

Figure 3: Illustration of the Semantic Gaussian Initialization process. The voxel-wise feature similarities are first computed between the voxel and TPV features to select top-k focal voxels as anchors. Then, a Gaussian projection layer is employed to generate Gaussian properties according to the semantics and positions of focal anchors.

are obtained with three spatial pooling layers to compress 3D features along the perpendicular axes. The compressed feature maps are further passed through 2D Convolution layers to refine global semantics:

$$\mathbf{F}_p = \text{Conv}_{\text{2D}}\left(\text{Pool}_{\text{p}}\left(\mathbf{F}_{\text{3D}}\right)\right), \quad p \in \{XY, YZ, ZX\} \tag{5}$$

where $\mathbf{F}_{XY} \in \mathbb{R}^{C\times X\times Y}, \mathbf{F}_{YZ} \in \mathbb{R}^{C\times Y\times Z}, \mathbf{F}_{ZX} \in \mathbb{R}^{C\times X\times Z}$ indicate the compressed features and the overall TPV representations are viewed as a list $\mathcal{F}_{\text{TPV}} = [\mathbf{F}_{XY}, \mathbf{F}_{YZ}, \mathbf{F}_{ZX}]$.

Focal Anchor Selection. To address spatial redundancy, we identify focal voxels with discriminative semantics as anchors for efficient and effective Gaussian initialization around occupied areas. Specifically, we first calculate the similarity score across local semantics embedded in $\mathbf{F}_{\text{voxel}}$ and global semantics embedded in $\mathcal{F}_{\text{TPV}}$:

$$M_{\text{sim}} = \text{sim}\left(\mathbf{F}_{\text{voxel}}, \mathcal{W} \odot \mathcal{F}_{\text{TPV}}\right) \tag{6}$$

where $M_{\text{sim}} \in \mathbb{R}^{X\times Y\times Z}$ represents the semantic similarity map, $\text{sim}(\cdot,\cdot)$ denotes the voxel-wise cosine similarity function, and $\mathcal{W}$ is the aggregation weight list for TPV representations. Then, we select top-k focal voxels with the highest similarity scores as anchors for Gaussian initializations:

$$\begin{aligned} \mathbf{P}_{\text{anc}} &= \text{top}_k(M_{\text{sim}}, K) \\ \mathbf{F}_{\text{anc}} &= (\mathbf{F}_{\text{voxel}} + \mathcal{W} \odot \mathcal{F}_{\text{TPV}})[\mathbf{P}_{\text{anc}}] \end{aligned} \tag{7}$$

where $\mathbf{P}_{\text{anc}} \in \mathbb{R}^{K\times 3}, \mathbf{F}_{\text{anc}} \in \mathbb{R}^{K\times C}$ represent the positions and semantic features of the selected K focal voxels respectively.

Semantic Gaussian Initialization. The focal voxels with high similarity scores possess consistent local and global semantics, thereby serving as ideal anchors for Gaussian initializations. As illustrated in Figure 3, the initialization process is formulated as follows:

$$[\mathbf{o}, \mathbf{s}, \mathbf{r}, \boldsymbol{\alpha}] = \text{MLP}_{\text{GS}}(\mathbf{F}_{\text{anc}}) \tag{8}$$

where offset $\mathbf{o}$, scale $\mathbf{s}$, rotation $\mathbf{r}$, and opacity $\boldsymbol{\alpha}$ are Gaussian properties projected from discriminative semantics of focal anchors.

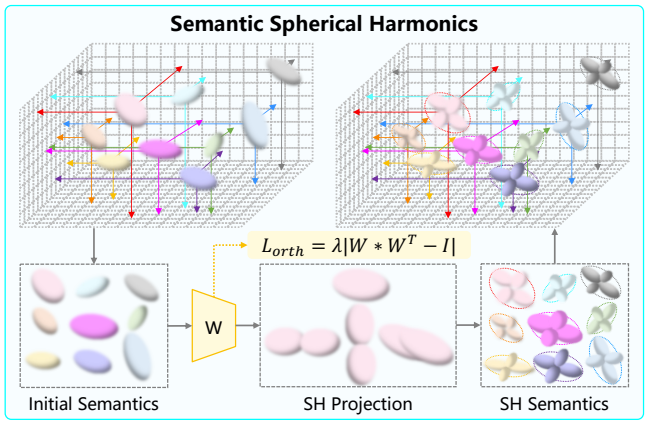

Figure 4: Illustration of the Semantic Spherical Harmonics. The initial Gaussian semantics are passed through SH projection to cope with SH coefficients, exploiting physical-aware contextual details. Furthermore, we employ an orthogonal loss on the projection matrix, considering that spherical harmonics form a complete set of orthogonal functions.

Then, the corresponding activation functions are applied for normalization:

$$\begin{aligned} \mathbf{m} &= \mathbf{P}_{\text{anc}} + \tanh(\mathbf{o}) \\ [\mathbf{s}, \boldsymbol{\alpha}] &= \text{sigmoid}([\mathbf{s}, \boldsymbol{\alpha}]) \\ \mathbf{r} &= \text{norm}\,(\mathbf{r}) \end{aligned} \tag{9}$$

By concatenating semantic features and normalized properties, the initialized Gaussians $\mathcal{G} = [\mathbf{F}_{\text{anc}}, \mathbf{m}, \mathbf{s}, \mathbf{r}, \boldsymbol{\alpha}] \in \mathbb{R}^{K\times(C+11)}$ efficiently cover focal regions with discriminative semantics.

3.3 Physical-aware Harmonics Enhancement

Design rationale. To generate realistic geometry structures without harming semantic accuracy, it is crucial to model physical-ware geometric details in accordance with voxel semantics. Therefore, the PHE module employs semantic spherical harmonics for modeling contextual details, and ensures semantic-geometry consistency with focal distribution alignment.

Semantic Spherical Harmonics. Spherical Harmonics (SH) demonstrates superior performance in RGB-based color rendering [13, 39], showcasing its powerful capability to model spatial variations and complex geometry. To leverage this capability for SSC, we employ the Semantic Spherical Harmonics (SSH), as shown in Figure 4, which incorporates semantic priors into SH basis functions:

$$S(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta, \phi) \tag{10}$$

where l, m are the degree and order controlling the level of details, (θ, ϕ) indicates the spherical coordinates, $Y_{lm}(\theta, \phi)$ denotes the SH basis function, and c_{lm} represents the expanded semantics. To cope with the Semantic Spherical Harmonics of L degrees, we employ 1D convolution layers to expand focal voxel semantics:

$$\{c_{lm}\} = \mathbf{F}_{\text{exp}} = Conv_{\text{1D}}(\mathbf{F}_{\text{anc}}) \tag{11}$$

where $\mathbf{F}_{\text{exp}} \in \mathbb{R}^{K\times 2^L C_N}$ corresponds to 2^L expanded semantics c_{lm} in the Semantic Spherical Harmonics. Considering that spherical harmonics form a complete set of orthogonal functions, to further enhance the capability of Semantic Spherical Harmonics in modeling contextual details, different degrees and orders (l, m) should correspond to distinct semantics. Therefore, we employ the soft orthogonality loss for supervision:

$$\mathcal{L}_{\text{orth}} = \lambda \left| W * W^{\text{T}} - I \right| \tag{12}$$

where W denotes the weight matrix in equation (11), I is the identity matrix, and λ is the regularization parameter. By promoting the orthogonality across projection weights, the expanded semantics $\{c_{lm}\}$ learn to capture distinct semantic information, thereby better modeling semantic-aware spatial variations and geometric details.

Focal Distribution Alignment. With the dual-branch 3D representations and sematic-guided Gaussian representations, we generate semantic scene completion predictions with task head and Gaussian superposition respectively:

$$\begin{aligned} \mathbf{V}_{\text{voxel}} &= \text{Head}\,(\mathbf{F}_{\text{voxel}} + \mathcal{W} \odot \mathcal{F}_{\text{TPV}}) \\ \mathbf{V}_{\text{gauss}}[\mathbf{x}] &= y(\mathbf{x}; [S_{\theta,\phi}(\mathbf{F}_{\text{exp}}), \mathbf{m}, \mathbf{s}, \mathbf{r}, \boldsymbol{\alpha}]) \end{aligned} \tag{13}$$

where $\mathbf{V}_{\text{voxel}}$ focuses on semantic accuracy, and $\mathbf{V}_{\text{gauss}}$ emphasizes geometry details. To promote the consistency between contextual semantics and geometry, we adopt an alignment loss across the predictions concerning focal voxels:

$$\begin{aligned} \mathcal{L}_{\text{align}} = &D_{\text{KL}}\left(\mathbf{V}_{\text{voxel}}[\mathbf{P}_{\text{anc}}] || \mathbf{V}_{\text{gauss}}[\mathbf{P}_{\text{anc}}]\right) \\ &+ D_{\text{KL}}\left(\mathbf{V}_{\text{gauss}}[\mathbf{P}_{\text{anc}}] || \mathbf{V}_{\text{voxel}}[\mathbf{P}_{\text{anc}}]\right) \end{aligned} \tag{14}$$

where $D_{\text{KL}}(\cdot||\cdot)$ is the Kullback-Leibler divergence measuring the difference between two distributions. The aligned predictions are then aggregated to generate the final prediction:

$$\mathbf{V}_{\text{SSC}} = \mathbf{V}_{\text{voxel}} + \mathbf{V}_{\text{gauss}} \tag{15}$$

3.4 Training Objective

Following [3, 26], we adopt the cross-entropy loss $\mathcal{L}_{\text{ce}}$, lovasz loss $\mathcal{L}_{\text{lovasz}}$ [2], and scene-class affinity loss $\mathcal{L}_{\text{scal}}$ to supervise the final predictions. The overall loss objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{lovasz}} + \mathcal{L}_{\text{scal}} + \mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{align}} \tag{16}$$

4 Experiments

4.1 Dataset and Evaluation Metrics

SemanticKITTI Benchmark [1] is constructed based on the KITTI Odometry Benchmark [5], comprising 22 sequences of real-world autonomous driving scenarios with 20 semantic categories. The semantic occupancy ground truth is generated as $256\times256\times32$ voxel grids with a resolution of $0.2m$, within the spatial range of $[0 \sim 51.2m, -25.6 \sim 25.6m, -2 \sim 4.4m]$. Following official protocols, sequences 00–07 and 09–10 are designated for training, sequence 08 for validation, and sequences 11–21 for testing.

SSCBench-KITTI-360 Benchmark [18] consists of 9 densely annotated sequences in urban driving environments, which are split into a training set with 8,487 frames from sequences (00, 02-05, 07), a validation set with 1,812 frames from sequence (06), and a test set with 2,566 frames from sequence (09). Semantic annotations across 19 classes cover a spatial region of $[0m, 51.2m]$ in the forward

Table 1: Camera-based 3D semantic scene completion results on the SemanticKITTI [1] validation set. † denotes the methods employing EfficientNetB7 [34] as image backbone, and * represents the methods utilizing ResNet50 [6] as image backbone. Best results are highlighted in bold, and second-best scores are underlined.

Method	SC IoU	SSC mIoU	car(3.92%)	bicycle(0.03%)	motorcycle(0.03%)	truck(0.16%)	other-veh.(0.20%)	person(0.07%)	bicyclist(0.07%)	motorcyclist(0.05%)	road(15.30%)	parking(1.12%)	sidewalk(11.13%)	other-grnd.(0.56%)	building(14.10%)	fence(3.90%)	vegetation(39.3%)	trunk(0.51%)	terrain(9.17%)	pole(0.29%)	traf.-sign(0.08%)
Mono-Input Methods																					
MonoScene† [3]	36.86	11.08	23.26	0.61	0.45	6.98	1.48	1.86	1.20	0.00	56.52	14.27	26.72	0.46	14.09	5.84	17.89	2.81	29.64	4.14	2.25
TPVFormer† [10]	35.61	11.36	23.81	0.36	0.05	8.08	4.35	0.51	0.89	0.00	56.50	20.60	25.87	0.85	13.88	5.94	16.92	2.26	30.38	3.14	1.52
OccFormer† [47]	36.50	13.46	25.09	0.81	1.19	25.53	8.52	2.78	2.82	0.00	58.85	19.61	26.88	0.31	14.40	5.61	19.63	3.93	32.62	4.26	2.86
IAMSSC* [41]	44.29	12.45	26.26	0.60	0.15	8.74	5.06	1.32	3.46	0.01	54.55	16.02	25.85	0.70	17.38	6.86	24.63	4.95	30.13	6.35	3.56
Temporal-Input Methods																					
VoxFormer* [19]	44.15	13.35	26.54	1.28	0.56	7.26	7.81	1.93	1.97	0.00	53.57	19.69	26.52	0.42	19.54	7.31	26.10	6.10	33.06	9.15	4.94
DepthSSC* [44]	45.84	13.28	25.94	0.35	1.16	6.02	7.50	2.58	**6.32**	0.00	55.38	18.76	27.04	0.92	19.23	8.46	26.37	4.52	30.19	7.42	4.09
Symphonies* [12]	41.92	14.89	28.68	2.54	**2.82**	**20.44**	13.89	**3.52**	2.24	0.00	56.37	15.28	27.58	0.95	21.64	8.40	25.72	6.60	30.87	9.57	5.76
HASSC* [36]	44.58	14.74	27.33	1.07	1.14	17.06	8.83	2.25	4.09	0.00	57.23	19.89	29.08	**1.26**	20.19	7.95	27.01	7.71	33.95	9.20	4.81
H2GFormer* [37]	44.69	14.29	28.21	0.95	0.91	6.80	9.32	1.15	0.10	0.00	57.00	21.74	29.37	0.34	20.51	7.98	27.44	7.80	36.26	9.88	5.81
CGFormer† [46]	45.99	16.87	34.32	**4.61**	2.71	19.44	7.67	2.38	4.08	0.00	**65.51**	20.82	32.31	0.16	23.52	9.20	26.93	8.83	**39.54**	10.67	7.84
SGN* [26]	46.21	15.32	33.31	0.61	0.46	6.03	9.84	0.47	0.10	0.00	59.10	19.05	29.41	0.33	25.17	**9.96**	28.93	**9.58**	38.12	**13.25**	7.32
SPHERE* (Ours)	**47.91**	**17.01**	**34.43**	0.57	0.84	14.66	**16.62**	1.02	0.85	0.00	60.32	**23.88**	**32.79**	0.11	**27.60**	9.71	**30.77**	9.43	38.23	13.18	**8.13**

direction, $[-25.6m, 25.6m]$ laterally, and $[-2.0m, 4.4m]$ vertically, which is uniformly discretized into $256 \times 256 \times 32$ voxel grids with a resolution of $0.2m$.

Evaluation Metrics. Following [26, 46], we adopt the Intersection over Union (IoU) of occupied voxels as the evaluation metric for the class-agnostic scene completion (SC). We also report the mean Intersection over Union (mIoU) with respect to all semantic classes to measure the performance of the semantic scene completion (SSC).

$$\begin{aligned} \text{IoU} &= \frac{TP}{TP + FP + FN} \\ \text{mIoU} &= \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \end{aligned} \tag{17}$$

where TP, FP, FN represent the number of true positive, false positive, and false negative occupancy predictions, and C stands for the total number of classes.

4.2 Implementation Details

Network Architecture. Following previous methods [19, 26, 36], we adopt the ResNet-50 [6] network with FPN [21] as the image encoder, generating 2D feature maps with 1/16 of the input resolution. The feature dimension of both 2D and 3D representations C is set to 128. The grid size of 3D feature volume is $(X, Y, Z) = (128, 128, 16)$, and the final SSC predictions are upsampled to the resolution of $256 \times 256 \times 32$. In SGI, we employ the dot production to compute voxel-wise feature similarity and select $K = 1024$ focal anchors for Gaussian initialization. In the PHE module, we set the regularization parameter for the soft orthogonality loss as $\lambda = 1e - 6$.

Training Setup. The RGB images from cam2 are cropped into size of 1220×370 as input for the SemanticKITTI dataset, and the RGB images from cam1 are resized to 1408×376 for the SSCBench-KITTI-360 dataset. We train SPHERE for 24 epochs on 4 NVIDIA A6000 GPUs, with a total batch size of 4. About 20 GB of GPU memory is consumed on each GPU while training. We employ a self-distillation training strategy [36] for more robust and efficient training process. The AdamW [24] optimizer is adopted with an initial learning rate of 2e-4 and a weight decay of 1e-2.

4.3 Quantitative Results

We conduct comparison experiments on two popular benchmarks, SemanticKITTI and SSCBench-KITTI-360, to evaluate the performance of our SPHERE against SOTA camera-based SSC methods.

SemanticKITTI: Table 1 presents the comparison results on the SemanticKITTI validation set. It can be observed that our SPHERE approach obtains superior performance of **47.91%** IoU for class-agnostic scene completion (SC) and **17.01%** mIoU for semantic scene completion (SSC), respectively. Compared to the best comparison method SGN [26] with ResNet50 backbone, our SPHERE achieves performance improvements of **1.70%** IoU and **1.69%** mIoU. Furthermore, compared with CGFormer [46], which employs EfficientNetB7 as backbone and Swin-T [23] as TPV feature backbone, our SPHERE still improves the SSC results by **1.92%** IoU and **0.14%** mIoU. The comparison results showcase the effectiveness of our SPHERE in generating more realistic SSC predictions through joint exploitation of semantic and physical information.

SSCBench-KITTI-360: Table 2 demonstrates the comparison results on the SSCBench-KITTI-360 test set, where our SPHERE approach outperforms all comparison methods with the performance of **48.59%** SC IoU and **20.56%** SSC mIoU, respectively. It is worth noting that Gaussian-based methods [9, 11] utilize a total of 38400 Gaussians to represent the scene, resulting in unsatisfying performance due to spatial redundancy. On the other hand, our SPHERE selects focal anchors for Gaussian initialization, achieving superior performance with only 1024 Gaussians around focal regions.

Table 2: Camera-based 3D semantic scene completion results on the SSCBench-KITTI-360 [1] test set. † denotes the methods employing EfficientNetB7 [34] as image backbone, and * represents the methods utilizing ResNet50 [6] as image backbone. Best results are highlighted in bold, and second-best scores are underlined.

Method	SC IoU	SSC mIoU	car(2.85%)	bicycle(0.01%)	motorcycle(0.01%)	truck(0.16%)	other-veh.(5.75%)	person(0.02%)	road(14.98%)	parking(2.31%)	sidewalk(6.43%)	other-grnd.(2.05%)	building(15.67%)	fence(0.96%)	vegetation(41.99%)	terrain(7.10%)	pole(0.22%)	traf.-sign(0.06%)	other-struct.(4.33%)	other-obj.(0.28%)
Mono-Input Methods																				
MonoScene† [3]	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
GaussianFormer* [11]	35.38	12.92	18.93	1.02	4.62	18.07	7.59	3.35	45.47	10.89	25.03	5.32	28.44	5.68	29.54	8.62	2.99	2.32	9.51	5.14
GaussianFormer2* [9]	38.31	13.90	21.08	2.55	4.21	12.41	5.73	1.59	54.12	11.04	32.31	3.34	32.01	4.98	28.94	17.33	3.57	5.48	5.88	3.54
TPVFormer† [10]	40.22	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
OccFormer† [47]	40.27	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
IAMSSC* [41]	41.80	12.97	18.53	2.45	1.76	5.12	3.92	3.09	47.55	10.56	28.35	4.12	31.53	6.28	29.17	15.24	8.29	7.01	6.35	4.19
Temporal-Input Methods																				
VoxFormer* [19]	38.76	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.18	4.97	28.99	14.69	6.51	6.92	3.79	2.43
DepthSSC* [44]	40.85	14.28	21.90	2.36	4.30	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.77	5.24	3.51
Symphonies* [12]	44.12	18.58	30.02	1.85	**5.90**	25.07	12.06	**8.20**	54.94	13.83	32.76	6.93	35.11	8.58	38.33	11.52	14.01	9.57	14.44	11.28
CGFormer† [46]	48.07	20.05	29.85	3.42	3.96	17.59	6.79	6.63	**63.85**	17.15	**40.72**	5.53	**42.73**	8.22	38.80	**24.94**	16.24	**17.45**	10.18	6.77
SGN* [26]	47.06	18.25	29.03	**3.43**	2.90	10.89	5.20	2.99	58.14	15.04	36.40	4.43	42.02	7.72	38.17	23.22	16.73	16.38	9.93	5.86
SPHERE* (Ours)	**48.59**	**20.56**	**33.08**	0.32	2.14	**25.62**	**12.79**	5.16	60.05	**17.21**	38.15	**9.23**	40.75	**9.24**	**41.97**	14.51	**18.23**	11.67	**17.22**	**12.84**

Table 3: Ablation study on the SemanticKITTI validation set for different architectural components of SPHERE.

SGI	PHE	IoU	mIoU
		44.15	13.35
✓		46.23	15.37
✓	✓	**47.91**	**17.01**

Table 4: Ablation study on the SemanticKITTI validation set validating the effectiveness of PHE.

SSH	FDA	IoU	mIoU
✓		47.18	16.53
	✓	47.37	16.10
✓	✓	**47.91**	**17.01**

Table 5: Efficiency evaluation against voxel-based method SGN [26] and Gaussian-based method GaussianFormer-2 [9].

Methods	Params (M)	Memory (G)	GFLOPs
GaussianFormer-2	71.57	39.21	1814.99
SGN	28.16	15.83	725.05
SPHERE (Ours)	28.79	16.13	841.59

4.4 Ablation Studies

To further investigate the effectiveness of our SPHERE approach and different components, we conduct ablation experiments on the SemanticKITTI validation set as follows:

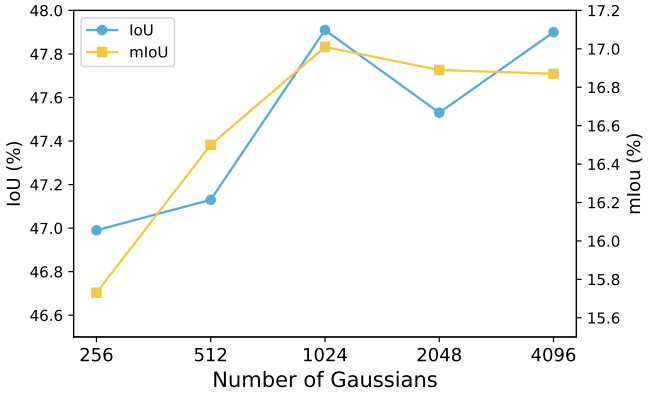

Figure 5: Effect of the number of Gaussians on the SSC performance for the SemanticKITTI validation set.

Ablation on Network Components. As shown in Table 3, we adopt VoxFormer [19] as the baseline method, which achieves 44.15% IoU and 13.5% mIoU. By adding SGI to the baseline, the model performances are improved by 2.08% IoU and 2.02% mIoU respectively, which showcases the effectiveness of leveraging Gaussian representations for improved geometric details. Furthermore, by incorporating PHE, SPHERE is able to model physical-aware contextual details with semantic spherical harmonics and promotes semantic-geometry consistency via focal distribution alignment. Therefore, the SSC performance is further enhanced with improvements of 3.90% IoU and 3.50% mIoU over the baseline method.

Ablation on Physical-aware Harmonic Enhancement. Tabel 4 presents the ablation experiments on SSH and FDA. It can be observed that incorporating SSH achieves higher performance improvements on the mIoU metric, since SSH focuses on capturing

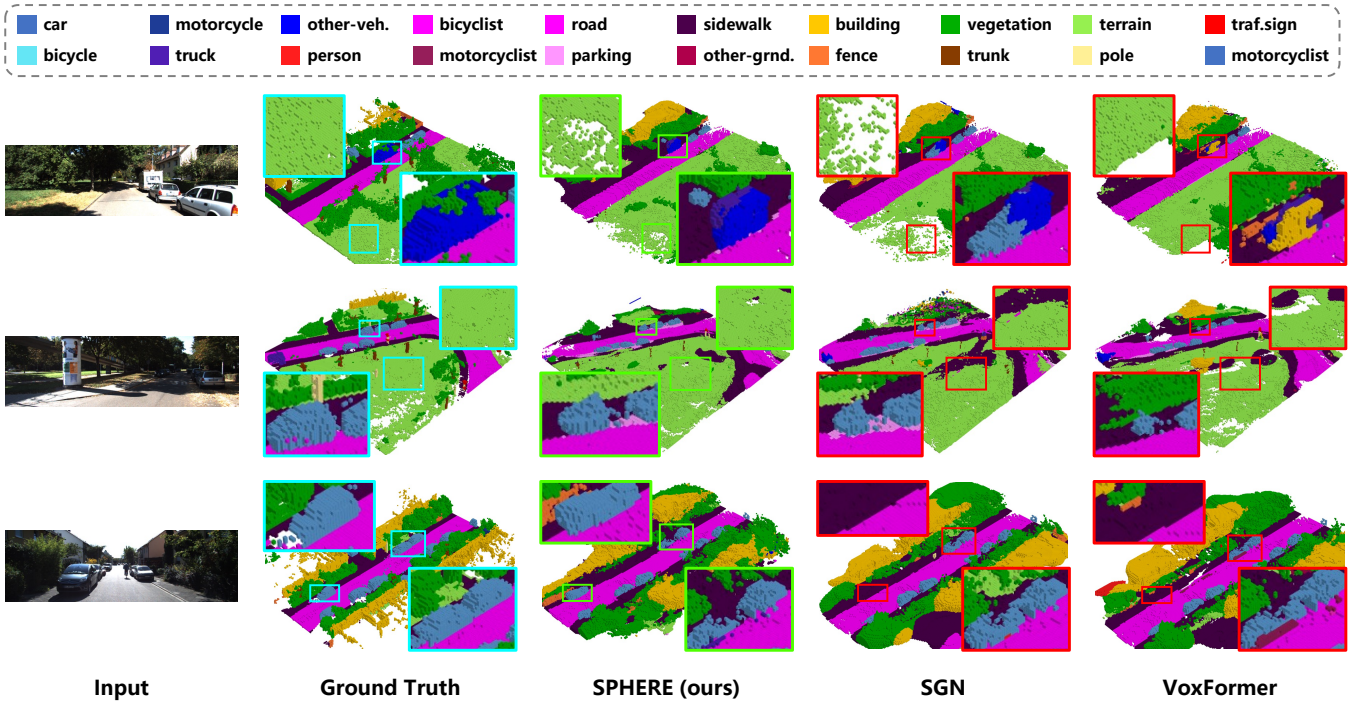

Figure 6: Qualitative visualization results on the SemanticKITTI validation set. Cyan boxes outline the occupancy ground truth. Red boxes indicate false occupancy predictions of the best comparison method SGN and baseline method VoxFormer, while green boxes indicate improved scene completion results generated by our SPHERE approach. Better viewed when zoomed in.

physical-aware contextual structures for more realistic details. On the other hand, integrating FDA better improves the IoU metric, since the aligned distribution promotes semantic-geometry consistency for more accurate predictions.

Efficiency Evaluation. Table 5 compares the training efficiency of our SPHERE against voxel-based and Gaussian-based methods. It can be observed that our SPHERE achieves a balance between memory for voxel representations and computation for Gaussian representations, generating accurate semantics and realistic geometry efficiently and effectively.

Effect of Gaussian numbers. Figure 5 illustrates the effect of the number of Gaussians K in SGI. The best performance is obtained with $K = 1024$, much fewer than that in Gaussian-based methods [9, 11]. This validates SPHERE alleviating the spatial redundancy issue with focal anchors for Gaussian initializations.

4.5 Qualitative Results

Figure 6 illustrates the visualization results on the SemanticKITTI validation set generated by VoxFormer [19], SGN [26], and our proposed SPHERE. We highlight the ground truth with cyan boxes as a reference for comparison. The red boxes indicate false predictions of SGN and VoxFormer, while the green boxes highlight improved results from our SPHERE approach. It can be observed that SPHERE generates improved SSC predictions with more accurate semantic predictions and more realistic geometric structures, especially for classes like car, other-vehicle, vegetation, and road. This improvement is attributed to the physical-aware modeling ability of Gaussian representations, and the semantic-geometry consistency promoted by focal distribution alignment.

5 Conclusion

In this work, we focus on a key challenge in camera-based SSC: voxel-based methods struggle to learn physical regularities for realistic details, while neural reconstruction methods yield confusing semantics due to spatial redundancy in autonomous driving scenes. To address this issue, we propose the SPHERE framework, which integrates voxel and Gaussian representations to jointly model semantic and physical information. Specifically, SGI leverages dual-branch representations to identify focal anchors for efficient and effective Gaussian initialization. Then, PHE incorporates the semantic spherical harmonics to model improved physical-aware details and ensures semantic-geometry consistency through focal distribution alignment between voxel and Gaussian representations. Extensive experiments on SemanticKITTI and SSCBench-KITTI-360 demonstrate the effectiveness of our SPHERE approach in achieving accurate semantics and realistic geometry.

Limitations. SPHERE leverages focal anchors with discriminative semantics for Gaussian initialization, taking advantage of both voxel and Gaussian representations for improved SSC performance. However, real-world scenarios involve complex conditions such as occlusion, motion blur, and low light, where low-quality semantic features damage model performance. In future work, we plan to incorporate explicit physical priors for improved performance with low-quality semantic features under complex input conditions.

Acknowledgments

This work was supported by the grants from Beijing Natural Science Foundation (L247006) and the National Natural Science Foundation of China (62525201, 62132001, 62432001).

References

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

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4413–4421.

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

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

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

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

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

[9] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. 2024. Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction. *arXiv preprint arXiv:2412.04384* (2024).

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

[11] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. 2024. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*. Springer, 376–393.

[12] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. 2024. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20258–20267.

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

[14] Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. 2024. Semcity: Semantic scene generation with triplane diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 28337–28347.

[15] Pengfei Li, Ruowen Zhao, Yongliang Shi, Hao Zhao, Jirui Yuan, Guyue Zhou, and Ya-Qin Zhang. 2023. Lode: Locally conditioned eikonal implicit scene completion from sparse lidar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8269–8276.

[16] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. 2023. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1486–1494.

[17] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. 2023. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1477–1485.

[18] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. 2023. SSCBench: A Large-Scale 3D Semantic Scene Completion Benchmark for Autonomous Driving. *arXiv preprint arXiv:2306.09001* (2023).

[19] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. 2023. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9087–9098.

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

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[22] Zhi Liu, Shaoyu Chen, Xiaojie Guo, Xinggang Wang, Tianheng Cheng, Hongmei Zhu, Qian Zhang, Wenyu Liu, and Yi Zhang. 2023. Vision-based uneven bev representation learning with polar rasterization and surface estimation. In *Conference on Robot Learning*. PMLR, 437–446.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

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

[25] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. 2024. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19936–19945.

[26] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Jongwon Ra, Yukai Ma, Laijian Li, and Yong Liu. 2024. Camera-based 3d semantic scene completion with sparse guidance network. *IEEE Transactions on Image Processing* (2024).

[27] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. 2023. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540* (2023).

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[29] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. 2024. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12404–12411.

[30] Jonah Philion and Sanja Fidler. 2020. 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*. Springer, 194–210.

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

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

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

[34] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[35] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. 2023. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8406–8415.

[36] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. 2024. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14792–14801.

[37] Yu Wang and Chao Tong. 2024. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5722–5730.

[38] Dongxu Wei, Zhiqi Li, and Peidong Liu. 2024. Omni-Scene: Omni-Gaussian Representation for Ego-Centric Sparse-View Scene Reconstruction. *arXiv preprint arXiv:2412.06273* (2024).

[39] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20310–20320.

[40] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. 2023. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17642–17651.

[41] Haihong Xiao, Hongbin Xu, Wenxiong Kang, and Yuqiong Li. 2024. Instance-aware monocular 3D semantic scene completion. *IEEE Transactions on Intelligent Transportation Systems* 25, 7 (2024), 6543–6554.

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

Artificial Intelligence, Vol. 35. 3101–3109.

[43] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 9421–9431.

[44] Jiawei Yao and Jusheng Zhang. 2023. Depthssc: Depth-spatial alignment and dynamic voxel resolution for monocular 3d semantic scene completion. *arXiv preprint arXiv:2311.17084* (2023).

[45] Zhaoda Ye, Xiangteng He, and Yuxin Peng. 2022. Unsupervised Cross-Media Hashing Learning via Knowledge Graph. *Chinese Journal of Electronics* 31, 6 (2022), 1081–1091.

[46] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. 2024. Context and geometry aware voxel transformer for semantic scene completion. *arXiv preprint arXiv:2405.13675* (2024).

[47] Yunpeng Zhang, Zheng Zhu, and Dalong Du. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9433–9443.

[48] Zhe Zhang, Bilin Wang, Zhezhou Yu, and Fengzhi Zhao. 2023. Attention guided enhancement network for weakly supervised semantic segmentation. *Chinese Journal of Electronics* 32, 4 (2023), 896–907.

[49] Jin Zheng, Botao Jiang, Wei Peng, and Qiaohui Zhang. 2024. Multi-Scale Binocular Stereo Matching Based on Semantic Association. *Chinese Journal of Electronics* 33, 4 (2024), 1010–1022.