SplatSSC: Decoupled Depth-Guided Gaussian Splatting for Semantic Scene Completion

Rui Qian[*1], **Haozhi Cao**[*1], **Tianchen Deng**[2], **Shenghai Yuan**[1], **Lihua Xie**[1]
[1]Nanyang Technological University
[2]Shanghai Jiao Tong University
{rqian003, haozhi002, shyuan, elhxie}@ntu.edu.sg, dengtiancheng@sjtu.edu.cn

Abstract

Monocular 3D Semantic Scene Completion (SSC) is a challenging yet promising task that aims to infer dense geometric and semantic descriptions of a scene from a single image. While recent object-centric paradigms significantly improve efficiency by leveraging flexible 3D Gaussian primitives, they still rely heavily on a large number of randomly initialized primitives, which inevitably leads to 1) inefficient primitive initialization and 2) outlier primitives that introduce erroneous artifacts. In this paper, we propose SplatSSC, a novel framework that resolves these limitations with a depth-guided initialization strategy and a principled Gaussian aggregator. Instead of random initialization, SplatSSC utilizes a dedicated depth branch composed of a Group-wise Multi-scale Fusion (GMF) module, which integrates multi-scale image and depth features to generate a sparse yet representative set of initial Gaussian primitives. To mitigate noise from outlier primitives, we develop the Decoupled Gaussian Aggregator (DGA), which enhances robustness by decomposing geometric and semantic predictions during the Gaussian-to-voxel splatting process. Complemented with a specialized Probability Scale Loss, our method achieves state-of-the-art performance on the Occ-ScanNet dataset, outperforming prior approaches by over 6.3% in IoU and 4.1% in mIoU, while reducing both latency and memory cost by more than 9.3%.

Code — https://github.com/Made-Gpt/SplatSSC

Introduction

3D scene understanding has garnered significant attention with the rapid evolution of embodied agents and autonomous driving (Li et al. 2024; Yi et al. 2025; Fusic and Sitharthan 2024; Zheng et al. 2025; Zammit and van Kampen 2023). As a key technology in this domain, 3D occupancy prediction (Tong et al. 2023; Huang et al. 2023; Wei et al. 2023; Tian et al. 2023; Wang et al. 2024b) and 3D Semantic Scene Completion (SSC) (Cao and de Charette 2022; Miao et al. 2023; Zhang, Zhu, and Du 2023; Li et al. 2023; Mei et al. 2024) have made remarkable progress. Early and conventional approaches for these tasks predominantly rely on grid-based representations. However, processing dense 3D volumes incurs prohibitive computational and memory costs. To mitigate this limitation, various efficiency-driven strategies have been explored, such as accelerating processing with Bird's-Eye-View (BEV) projections (Yu et al. 2023; Hou et al. 2024), or leveraging the natural sparsity of scenes with sparse voxels (Tang et al. 2024; Li et al. 2023) and points (Shi et al. 2024; Wang et al. 2024a).

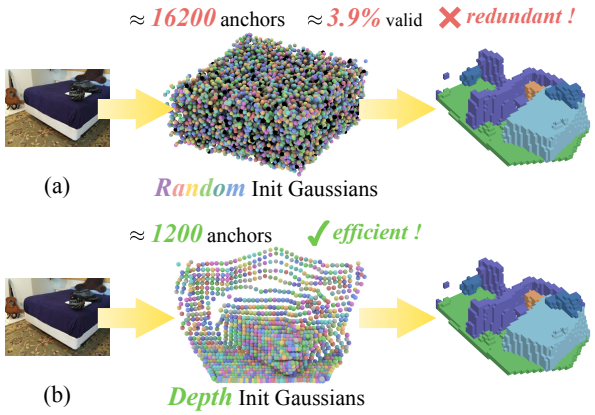

Figure 1: Comparison with prior framework. (a) Recent transformer-based SSC frameworks start with a large set of randomly initialized Gaussian primitives, introducing redundancy. (b) Our framework starts with a compact yet targeted set of Gaussian primitives, guided by geometric priors.

Despite the devoted efforts, such methods remain inherently constrained by their discrete and grid-aligned nature, which struggle to model the sparse geometry efficiently. A recent paradigm shift towards object-centric representations, pioneered by GaussianFormer (Huang et al. 2024), has achieved a breakthrough. By utilizing flexible 3D Gaussian primitives (Kerbl et al. 2023) to represent the scene, this approach strikes a new balance between performance and efficiency. Building upon this foundation, subsequent works (Huang et al. 2025) have advanced this field by developing more principled aggregation methods based on Gaussian Mixture Models (GMMs) and adapting the paradigm to indoor scenes for incremental perception (Wu et al. 2025; Zhang et al. 2025; Wang et al. 2025a).

While the object-centric paradigm offers a promising di-

[*]These authors contributed equally.

rection, its application in vision-only settings faces a foundational challenge: ***how to efficiently initialize and reliably supervise 3D primitives using only monocular cues***. To ensure complete coverage of the target 3D space without geometric cues, the predominant strategy is to randomly distribute numerous primitives throughout the 3D volume, as shown in Figure 1(a). This leads to two critical, coupled limitations: 1) *Inefficient Primitive Initialization.* A significant portion of the model's capacity is inevitably wasted on representing empty or unknown space, making the random distribution strategy inherently redundant. 2) *Fragile Aggregation of Outliers.* Existing Gaussian-to-voxel splatting strategies (Huang et al. 2024, 2025) lack an effective rejection mechanism to mitigate the impact of outlier primitives. This allows outliers to spurious semantics on distant voxels, creating "floaters" in otherwise empty space.

To this end, we introduce SplatSSC, a novel framework designed to tackle inefficient initialization and fragile aggregation in object-centric SSC. Rather than starting with a large set of random primitives, SplatSSC leverages geometric priors to guide the primitive initialization, as shown in Figure 1(b), reducing redundancy while maintaining representational capacity. We generate this prior using a tailored depth branch, equipped with our proposed *Group-wise Multi-scale Fusion* (GMF) module. GMF integrates multi-scale image features and depth features from a pretrained depth estimator via *Group Cross-Attention* (GCA) for efficient multi-modal fusion. The resulting geometric priors subsequently guide a lifter to initialize a sparse yet targeted set of Gaussian primitives, which are then refined through a standard multi-stage encoder. To address the "floaters" that plague existing aggregators when dealing with sparse outliers, we propose the *Decoupled Gaussian Aggregator* (DGA), which renders the final semantic grid by completely decomposing semantic and geometry prediction robustly. Furthermore, to ensure stable geometric learning, we design the specialized *Probability Scale Loss* to apply soft and progressive supervision to the intermediate encoder layers.

In summary, our contributions are as follows:

- We propose an efficient object-centric paradigm for monocular SSC, namely SplatSSC, which features a depth-guided strategy for initializing a sparse and targeted set of Gaussian primitives.
- We introduce the *Group-wise Multi-scale Fusion (GMF)* module with a *Group Cross-Attention (GCA)* core to efficiently generate a high-quality geometric prior.
- We design the *Decoupled Gaussian Aggregator (DGA)* that decouples geometry and semantics to eliminate aggregation artifacts from sparse primitives robustly.
- We propose a *Probability Scale Loss* to provide auxiliary geometric supervision for robust end-to-end training.

Related Work

3D Semantic Scene Completion. 3D Semantic Scene Completion (SSC) infers dense geometry and semantics from partial observations. Early approaches (Song et al. 2017; Zhang et al. 2019; Wang et al. 2019) primarily focused on indoor scenes using depth-only input, where deep convolutional networks (CNNs) and Truncated Signed Distance Function (TSDF) representations were widely employed. To improve semantic understanding, subsequent methods (Li et al. 2019, 2020; Wang et al. 2023) fuse features from both RGB and depth inputs. In parallel, LiDAR-based SSC approaches (Roldão, de Charette, and Verroust-Blondet 2020; Yan et al. 2021; Yang et al. 2021) have been developed for autonomous driving and also rely on CNN architectures.

A recent trend has shifted towards vision-only methods. MonoScene (Cao and de Charette 2022) pioneered this direction using a dense 2D-to-3D lifting with UNet architecture (Ronneberger, Fischer, and Brox 2015), but this approach suffered from inherent depth ambiguity. To address this, OccDepth (Miao et al. 2023) and ISO (Yu et al. 2024a) introduced depth-aware strategies by leveraging stereo depth and pretrained depth networks, respectively. Concurrently, to tackle the inefficiency of dense voxel processing, VoxFormer (Li et al. 2023), a two-stage model, proposed a sparse-to-dense Transformer method based on generating proposals from a geometry prior. Subsequent works continue to advance this paradigm (Mei et al. 2024; Yu et al. 2024b; Jiang et al. 2024), focusing on unified pipelines, context-aware modeling, and instance-level reasoning.

While these Transformers improve accuracy, they remain bound to grid-aligned voxel queries. Our work takes a different route through a flexible object-centric formulation inspired by (Wu et al. 2025; Huang et al. 2025).

Object-centric 3D Scene Representation. A recent paradigm shift in occupancy prediction, pioneered by GaussianFormer (Huang et al. 2024), moves beyond grid-aligned queries to object-centric representation using 3D Gaussian primitives. This approach leverages the inherent sparsity of 3D scenes by representing them as a collection of continuous ellipsoids, which are then rendered into a dense semantic grid via an efficient Gaussian-to-voxel splatting mechanism. This marked a significant departure from discrete, voxel-based frameworks (Li et al. 2023; Tang et al. 2024).

Sequential works (Huang et al. 2025; Zhao et al. 2025) further advanced this paradigm by introducing a principled probabilistic framework via GMMs and incorporating LiDAR-guided initialization to replace random placement. In parallel, EmbodiedOcc (Wu et al. 2025) first adapted this object-centric paradigm to the unique challenges of indoor perception. It focuses on online incremental scene understanding, where confidence refinement is applied to continuously update the Gaussian representation as an agent explores the environment. Following this, RoboOcc and EmbodiedOcc++ (Zhang et al. 2025; Wang et al. 2025a) extended this paradigm through geometry-aware refinement, leveraging opacity cues and planar constraints to enhance stability and structural fidelity.

However, object-centric approaches widely employ random Gaussian primitive initialization, which introduces significant redundancy, as most primitives are used to represent empty space. In contrast, our method directly tackles this problem by leveraging a depth prior to generate a compact but more targeted set of primitives.

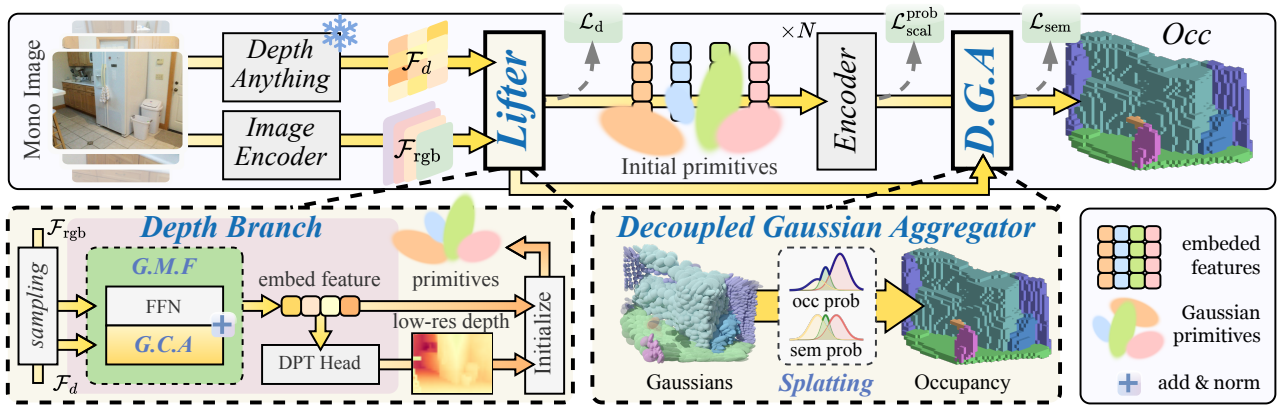

Figure 2: An overview of our proposed SplatSSC architecture. Given a single input image, our model employs two parallel branches: a trainable image encoder to extract multi-scale image features, and a frozen, pretrained *Depth-Anything* model to extract depth features. After a sampling step, both features are fed into the proposed *Group-wise Multi-scale Fusion* (GMF) block and a two-convolution layer depth head, yielding a refined feature map and a low-resolution depth map. These outputs are then lifted to initialize a set of 3D Gaussian primitives. Subsequently, the primitives are processed by a multi-stage encoder and finally passed to our *Decoupled Gaussian Aggregator* (DGA) to render the final semantic voxels.

Methodology

Problem Setup

Formally, given a single input RGB image $\mathcal{I}_{\text{rgb}}$, the local prediction task is to infer the dense semantic voxel grid $\mathcal{V}_{\text{loc}}$ and the underlying set of sparse Gaussian primitives $\mathcal{G}_{\text{loc}}$ that represent the scene within the current camera frustum. This process is defined as:

$$(\mathcal{V}_{\text{loc}}, \mathcal{G}_{\text{loc}}) = \mathcal{M}_{\text{loc}}(\mathcal{I}_{\text{rgb}}), \tag{1}$$

where $\mathcal{M}_{\text{loc}}$ is our prediction model. The output grid $\mathcal{V}_{\text{loc}} \in \{0, 1, ..., C-1\}^{X_{\text{loc}} \times Y_{\text{loc}} \times Z_{\text{loc}}}$ assigns each voxel a label from C semantic classes, with class 0 denoting empty space. The scene itself is represented by the set of N refined Gaussian primitives $\mathcal{G}_{\text{loc}} = \{G_i\}_{i=1}^{N}$. Each primitive G_i is parameterized by its geometric and semantic properties: a mean $\boldsymbol{\mu}_i \in \mathbb{R}^3$, a scale vector $\mathbf{s}_i \in \mathbb{R}^3$, a rotation quaternion $\mathbf{q}_i \in \mathbb{R}^4$, an opacity $\mathbf{a}_i \in [0, 1]$, and a semantic logit vector $\mathbf{c}_i \in \mathbb{R}^{C-1}$. The scale and rotation are used to construct the full anisotropic covariance matrix $\boldsymbol{\Sigma}_i$:

$$\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^T \mathbf{R}_i^T, \ \mathbf{S}_i = \text{diag}(\mathbf{s}_i), \ \mathbf{R}_i = \text{q2r}(\mathbf{q}_i). \tag{2}$$

where $\text{q2r}(\cdot)$ converts a quaternion into a rotation matrix and $\text{diag}(\cdot)$ forms a diagonal scaling matrix.

Overview

The architecture of our approach is illustrated in Figure 2. We first process the input image $\mathcal{I}_{\text{rgb}}$ with an image encoder, composed of a lightweight image backbone EfficientNet (Tan and Le 2019) and FPN (Lin et al. 2017), to extract multi-scale image features $\mathcal{F}_{\text{rgb}} = \{f_{\text{rgb}}^l\}_{l=1}^{L}$, where L is the scale number. Simultaneously, a pretrained depth estimation model *Depth-Anything* (Yang et al. 2024) is employed to produce powerful depth features $\mathcal{F}_d$. These two feature streams are then fed into our specialized *depth branch*, which employs the proposed GMF module to produce the fused depth features $\mathcal{F}_d'$ and the refined depth map $\mathcal{I}_d$. The resulting $\mathcal{F}_d'$ and $\mathcal{I}_d$ are then fed to a lifting module to obtain the initial Gaussian primitives $\mathcal{G}_o$ with good geometry prior. Subsequently, $\mathcal{G}_o$ is refined by a series of encoder blocks cyclically, following EmbodiedOcc. Given the refined primitives, the 3D semantic voxels are obtained by our DGA $\hat{\mathcal{V}}_{\text{agg}}$. By first leveraging the depth branch to generate a highly compact set of primitives with geometrically grounded initial locations, we tackle the inefficiency inherent in random initialization strategies. Subsequently, our DGA transforms primitives into semantic voxels, overcoming the fragility of prior aggregation methods. This enables our framework to achieve state-of-the-art (SOTA) performance while maintaining high efficiency with significantly fewer primitives.

Depth Branch

While recent monocular 3D completion methods (Wu et al. 2025; Yu et al. 2024a) leverage pretrained depth estimators, they tend to utilize depth information as a secondary guiding signal: either refining geometric distributions or informing feature learning. However, this approach neglects the rich latent features generated by depth networks. In contrast, our framework proposes a dual-pronged strategy: we use the depth map as a direct geometric prior, while simultaneously employing the latent depth features as the initial embeddings for 3D primitives. This not only ensures primitives are grounded in both geometry (*where*) and semantics (*what*), but necessitates a more advanced fusion mechanism. To fulfill this demand, we design a dedicated depth branch. Inspired by prior works (Ma et al. 2020; Jia et al. 2025), this branch fuses multi-scale image features and depth cues via our GMF mechanism. Specifically, GMF is a Transformer-like block comprising the proposed GCA layer followed by a point-wise FFN (Vaswani et al. 2017). The resulting fused features $\mathcal{F}_d'$ are then processed by two convolutional layers to produce the refined depth map $\mathcal{I}_d$.

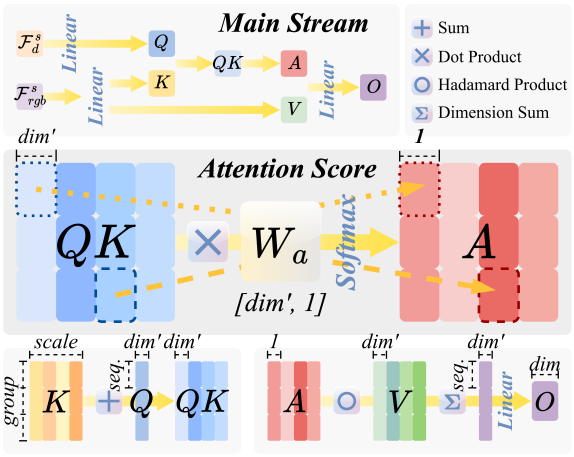

Figure 3: Illustration of the proposed GCA layer. The weight matrix W_a is shared across different groups and scales, thus reducing memory consumption and computational cost.

Group Cross-Attention. The architecture of our GCA module is illustrated in Figure 3. The process begins by sampling features from the input depth features $\mathcal{F}_d$ and multi-scale image features $\mathcal{F}_{\text{rgb}}$, using a set of predefined reference points normalized to the $[0, 1]$ range. This step yields the sampled features, denoted as $\mathcal{F}_d^s$ and $\mathcal{F}_{\text{rgb}}^s = \{f_{\text{rgb}}^{s,l}\}_{l=1}^L$ respectively. To balance performance and efficiency, we split these features into G groups along the channel dimension, where each group has a reduced feature dimensionality of $D_g = D/G$. The Query Q_g is projected from sampled depth features, while the Key K_g^l and Value V_g^l are projected from sampled image features at each scale l:

$$Q_g = (\mathcal{F}_d^s W_q)^g, K_g^l = (f_{\text{rgb}}^{s,l} W_k)^g, V_g^l = (f_{\text{rgb}}^{s,l} W_v)^g, \quad (3)$$

where W_q, W_k, and W_v are linear projection matrices for Query, Key, and Value, respectively. $l \in \{1, ..., L\}$ denotes the scale index. Inspired by the efficient design of Deformable Attention (Zhu et al. 2021), we adopt a lightweight linear projection mechanism in place of the standard dot-product attention. To elaborate, the attention scores are computed by feeding the element-wise sum of queries and keys into a shared projection $W_a \in \mathbb{R}^{D_g \times 1}$:

$$A_g^l = \mathbb{S}_l \left(W_a (Q_g + K_g^l) \right), \quad (4)$$

where $\mathbb{S}_l(\cdot)$ denotes the Softmax operation across the scale dimension, and g indexes feature groups. With the group-wise formulation, both scale-wise attention and projection are computed within each group, allowing W_a to be shared across different groups and scales. This design significantly reduces parameter overhead and computation.

The final fused representation is obtained by aggregating value features V_g^l using Hadamard product $\circ$ with the attention scores, followed by group concatenation $\mathbb{C}_g(\cdot)$ and a linear projection W_o:

$$\mathcal{F}_d' = \mathbb{C}_g \left(\sum_{l=1}^{L} A_g^l \circ V_g^l \right) W_o. \quad (5)$$

Efficiency Analysis. The design of GCA is computationally lean. Standard cross-attention has a complexity of $\mathcal{O}(LN^2D)$, where N is the sequence length. In contrast, by employing a group-wise mechanism and replacing the quadratic-cost dot-product with a linear-cost MLP, GCA significantly reduces the complexity. The dominant cost of our module becomes $\mathcal{O}(ND^2(L+2)/G)$, which is substantially more efficient, especially for long sequences.

Decoupled Gaussian Aggregator

Gaussian-to-voxel splatting is a critical step for object-centric approaches, which dictates the final quality of the occupancy output. While GaussianFormer first enabled object-centric aggregation, its additive nature leads to redundancy. The subsequent *Probabilistic Gaussian Superposition* (PGS) model proposed in GaussianFormer-2, though theoretically elegant, introduces a flawed decoupling of geometry and semantics and therefore falls short when tackling outlier primitives. To address these limitations, we propose the DGA, a novel strategy that reformulates the task into two distinct prediction pathways: *Geometry Occupancy Prediction* and *Conditional Semantic Distribution*.

Analysis of Probabilistic Gaussian Superposition. The PGS models the semantic occupancy prediction at a point $\mathbf{x}$ as a two-part process: a geometric occupancy probability $\alpha(\mathbf{x})$ and a conditional semantic expectation $\mathbf{e}(\mathbf{x}; \mathcal{G})$:

$$\alpha(x) = 1 - \prod_{i \in \mathcal{N}(\mathbf{x})} \left(1 - \alpha(\mathbf{x}; G_i)\right), \quad (6)$$

$$\mathbf{e}(\mathbf{x}; \mathcal{G}) = \sum_{i=1}^{N} p(G_i|\mathbf{x}) \tilde{\mathbf{c}}_i = \frac{\sum_{i=1}^{N} p(\mathbf{x}|G_i) \mathbf{a}_i \tilde{\mathbf{c}}_i}{\sum_{j=1}^{N} p(\mathbf{x}|G_j) \mathbf{a}_j}, \quad (7)$$

$$p(\mathbf{x}|G_i) = \frac{1}{(2\pi)^{3/2} |\mathbf{\Sigma}_i|^{1/2}} \alpha(\mathbf{x}; G_i), \quad (8)$$

where $p(\mathbf{x} \mid G_i)$ denotes the conditional Gaussian probability of point $\mathbf{x}$ under the ith primitive, and $p(G_i \mid \mathbf{x})$ represents the normalized posterior of selecting G_i for $\mathbf{x}$ under a uniform prior. $\alpha(\mathbf{x}; G_i) = \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i))$ is the un-normalized Gaussian kernel. The key flaw in this formulation lies in how the learned opacity $\mathbf{a}_i$ is used. While intended to represent a primitive's existence confidence, it is instead employed as the prior probability in the GMM. The negative consequence of this choice becomes evident when considering an isolated outlier primitive G_n. For any point $\mathbf{x}^f$ in its immediate vicinity, the likelihood $p(\mathbf{x}^f|G_m)$ for all other distant primitives $G_{m \neq n}$ approaches zero. This causes the normalization term in the posterior calculation to be dominated by the outlier itself. Hence, the posterior probability $p(G_n|\mathbf{x}^f)$ collapses to unity, regardless of the effect of the low-confidence prior $\mathbf{a}_n$:

$$\begin{aligned} p(G_n|\mathbf{x}^f) &= \frac{p(\mathbf{x}^f|G_n)\mathbf{a}_n}{\sum_{j=1}^{N} p(\mathbf{x}^f|G_j)\mathbf{a}_j} \\ &\approx \frac{p(\mathbf{x}^f|G_n)\mathbf{a}_n}{p(\mathbf{x}^f|G_n)\mathbf{a}_n + 0} = 1. \end{aligned} \quad (9)$$

Accordingly, the semantic expectation at this point reduces to $\mathbf{e}(\mathbf{x}^f; \mathcal{G}) \approx \tilde{\mathbf{c}}_n$, with the learned opacity $\mathbf{a}_n$ nullified by the posterior normalization.

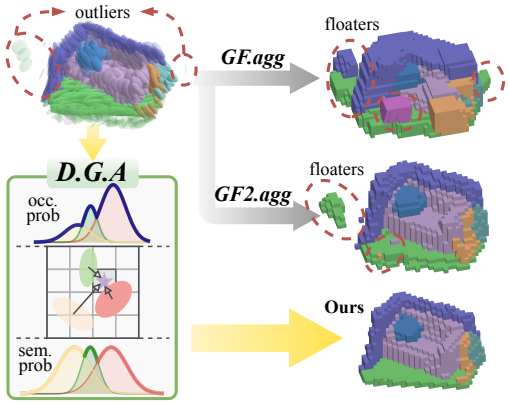

Figure 4: Illustration of the proposed DGA. While *GF.agg* (Huang et al. 2024) and *GF2.agg* (Huang et al. 2025) wrongly produces the "floaters" from outliers, our DGA remains robust, as the low occupancy probability directly suppresses its erroneous semantic contribution.

This issue is further exacerbated when considering the geometry prediction, where the opacity $\mathbf{a}_i$ is decomposed and depends solely on the Gaussian kernel. As such, even a low-confidence outlier can yield a high occupancy value for nearby points. Consequently, the voxel $\mathbf{x}^f$ is likely to be incorrectly activated as occupied by the semantic label of the outlier G_n, producing the characteristic "floaters".

Geometric Occupancy Prediction. Due to the exponential decay of the Gaussian kernel, only primitives in the local vicinity of $\mathbf{x}$ have a meaningful influence. Therefore, we only consider contributions from a neighborhood of relevant Gaussian primitives for efficiency, denoted as $\mathcal{N}(\mathbf{x})$. The occupancy is then modeled as a probabilistic OR operation over this local set. Crucially, each primitive's influence is modulated by its learned opacity $\mathbf{a}_i$, which we interpret as its existence confidence. This explicit use of opacity is a key difference from PGS:

$$\alpha'(x) = 1 - \prod_{i \in \mathcal{N}(\mathbf{x})} (1 - \alpha(\mathbf{x}; G_i) \cdot \mathbf{a}_i)\,. \tag{10}$$

This natural gating mechanism suppresses the influence of low-confidence outliers on the final occupancy probability.

Conditional Semantic Distribution. Concurrently, we predict the conditional semantic distribution $e(\mathbf{x})$ under the assumption that the position $\mathbf{x}$ is occupied. This is achieved by using GMM, where we leverage the normalized semantic weights of each Gaussian component. This design decouples the semantic prediction from the opacity parameter $\mathbf{a}_i$, forcing the model to rely solely on the geometric proximity and the learned softmax-normalized semantic properties $\tilde{\mathbf{c}}_i$ of each primitive. The posterior probability for each semantic class k is then computed as:

$$e^k(\mathbf{x}) = \sum_{i \in \mathcal{N}(\mathbf{x})} p(G_i|\mathbf{x}) = \frac{\sum_{i \in \mathcal{N}(\mathbf{x})} p(\mathbf{x}|G_i) \cdot \tilde{\mathbf{c}}_i^k}{\sum_{j \in \mathcal{N}(\mathbf{x})} p(\mathbf{x}|G_j)}. \tag{11}$$

Probabilistic Fusion. Finally, the two decoupled pathways are fused to compute the final probability distribution $\hat{\mathbf{y}}_x$ for each 3D position $\mathbf{x}$. The probabilities for each valid semantic class k and the empty class are defined as:

$$\begin{cases} \hat{\mathbf{y}}_x^k = \alpha'(\mathbf{x}) \cdot e^k(\mathbf{x}) \\ \hat{\mathbf{y}}_x^{\text{empty}} = 1 - \alpha'(\mathbf{x}) \end{cases}. \tag{12}$$

This formulation serves as a principled and fully differentiable gating mechanism. A low occupancy probability $\alpha'(\mathbf{x})$, often resulting from an outlier primitive, directly suppresses any erroneous semantic prediction $e^k(\mathbf{x})$, thus elegantly eliminating "floaters" without complex heuristics. We demonstrate this effect in Figure 4.

Training Objective

Our model is trained via a two-stage strategy, where the first stage establishes a robust geometric prior before training the full network end-to-end. Throughout both stages, the pre-trained model *Depth-Anything-V2* is kept frozen.

Stage 1: Depth Branch Pre-training. In this stage, we exclusively train our depth branch to produce a high-quality geometry prior. This module is supervised by a composite depth loss $\mathcal{L}_{\text{d}}$, similar with prior works (Laina et al. 2016; Wang et al. 2025b):

$$\mathcal{L}_{\text{d}} = \lambda_1 \mathcal{L}_{\text{huber}}^{\text{depth}} + \lambda_2 \mathcal{L}_{\text{huber}}^{\text{pts}} + \lambda_3 \mathcal{L}_{\text{grad}}, \tag{13}$$

where the terms are the depth Huber loss, point cloud Huber loss, and gradient matching Huber losses.

Stage 2: End-to-End SplatSSC Training. In this stage, we train the entire SplatSSC network. To prevent the model from being overly constrained by the initial depth predictions, while maintaining a robust geometric prior, we remove $\mathcal{L}_{\text{d}}$ and introduce our proposed *Probability Scale Loss* $\mathcal{L}_{\text{scal}}^{\text{prob}}$ as a soft geometric supervision. The training objective is therefore optimized with a final composite loss $\mathcal{L}_{\text{ssc}}$:

$$\mathcal{L}_{\text{ssc}} = \mathcal{L}_{\text{sem}} + \lambda_4 \mathcal{L}_{\text{scal}}^{\text{prob}}, \tag{14}$$

where $\mathcal{L}_{\text{sem}} = \lambda_5 \mathcal{L}_{\text{focal}} + \lambda_6 \mathcal{L}_{\text{lovasz}}$ is the primary semantic segmentation loss adopted by EmbodiedOcc. Our loss $\mathcal{L}_{\text{scal}}^{\text{prob}}$ extends the geometry-aware scale loss $\mathcal{L}_{\text{scal}}^{\text{geo}}$ from MonoScene (Cao and de Charette 2022), adapting it to supervise the predicted occupancy probability across all n encoder layers. To account for the progressive refinement across stages, we introduce a linear weighting schedule, which imposes weaker constraints on early-stage predictions and gradually enforces stronger consistency at deeper layers:

$$\mathcal{L}_{\text{scal}}^{\text{prob}} = \frac{1}{2} \sum_{i=1}^{n-1} \frac{i}{n} \cdot \mathcal{L}_{\text{scal}}^{\text{geo},i} + \mathcal{L}_{\text{scal}}^{\text{geo},n}, \tag{15}$$

where i is the layer index. The loss weights are set as $\lambda_1 = 10$, $\lambda_2 = 20$, $\lambda_3 = \lambda_4 = 0.5$, $\lambda_5 = 100$, and $\lambda_6 = 2$.

Experiments

To evaluate the effectiveness of our SplatSSC, we conduct extensive experiments on the high-quality indoor datasets Occ-ScanNet and Occ-ScanNet-mini (Yu et al. 2024a). Details about datasets, implementation, and evaluation metrics are included in our supplementary material from our code.

Dataset	Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
Occ-ScanNet	TPVFormer	$\mathcal{I}_{\text{rgb}}$	33.39	6.96	32.97	14.41	9.10	24.01	41.49	45.44	28.61	10.66	35.37	25.31	24.94
	GaussianFormer	$\mathcal{I}_{\text{rgb}}$	40.91	20.70	42.00	23.40	17.40	27.00	44.30	44.80	32.70	15.30	36.70	25.00	29.93
	MonoScene	$\mathcal{I}_{\text{rgb}}$	41.60	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	24.62
	ISO	$\mathcal{I}_{\text{rgb}}$	42.16	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61	28.71
	SurroundOcc	$\mathcal{I}_{\text{rgb}}$	42.52	18.90	49.30	24.80	18.00	26.80	42.00	44.10	32.90	18.60	36.80	26.90	30.83
	EmbodiedOcc	$\mathcal{I}_{\text{rgb}}$	53.95	40.90	50.80	41.90	33.00	41.20	55.20	61.90	43.80	35.40	53.50	42.90	45.48
	EmbodiedOcc++	$\mathcal{I}_{\text{rgb}}$	54.90	36.40	53.10	41.80	34.40	42.90	57.30	64.10	45.20	34.80	54.20	44.10	46.20
	RoboOcc	$\mathcal{I}_{\text{rgb}}$	56.48	45.36	53.49	44.35	34.81	43.38	56.93	63.35	46.35	36.12	55.48	44.78	47.67
	SplatSSC (Ours)	$\mathcal{I}_{\text{rgb}}$	**62.83**	**49.10**	**59.00**	**48.30**	**38.80**	**47.40**	**62.40**	**67.00**	**49.50**	**42.60**	**60.70**	**45.40**	**51.83**
Occ-ScanNet-mini	MonoScene	$\mathcal{I}_{\text{rgb}}$	41.90	17.00	46.20	23.90	12.70	27.00	29.10	34.80	29.10	9.70	34.50	20.40	25.90
	ISO	$\mathcal{I}_{\text{rgb}}$	42.90	21.10	42.70	24.60	15.10	30.80	41.00	43.30	32.20	12.10	35.90	25.10	29.40
	EmbodiedOcc	$\mathcal{I}_{\text{rgb}}$	55.13	29.50	49.40	41.70	36.30	41.90	60.40	59.60	46.30	34.50	58.00	43.50	45.57
	EmbodiedOcc++	$\mathcal{I}_{\text{rgb}}$	55.70	23.30	51.00	42.80	39.30	43.50	**65.60**	**64.00**	**50.70**	**40.70**	60.30	**48.90**	48.20
	SplatSSC (Ours)	$\mathcal{I}_{\text{rgb}}$	**61.47**	**36.60**	**55.70**	**46.50**	**40.10**	**45.60**	64.50	62.40	48.60	30.60	**61.20**	45.39	**48.87**

Table 1: Local Prediction Performance on the Occ-ScanNet dataset. The best results are highlighted in **bold**, while the second-best are underlined.

Number	Scale Range	Mem.↓ (MiB)	Time↓ (ms)	Train	IoU	mIoU
19200	[0.01, 0.08]	3.122	135.18	✓	**62.77**	47.69
19200	[0.01, 0.16]	4.978	134.25	✓	60.64	43.31
19200	[0.01, 0.32]	14.380	134.51	OOM	/	/
4800	[0.01, 0.08]	3.158	123.27	✓	62.23	47.20
4800	[0.01, 0.16]	3.108	122.63	✓	61.53	46.74
4800	[0.01, 0.32]	5.854	122.70	✓	60.78	46.96
1200	[0.01, 0.08]	3.104	116.20	✓	60.18	48.32
1200	[0.01, 0.16]	3.112	115.56	✓	61.47	**48.87**
1200	[0.01, 0.32]	3.126	114.75	✓	57.09	42.38

Table 2: Ablation on Gaussian Parameters. Memory (Mem.) usage and time are measured on one 3090 GPU.

GMF	GF.agg	GF2.agg	DGA	IoU	mIoU
-	✓	-	-	11.64	12.62
-	-	✓	-	27.54	17.27
-	-	-	✓	48.85	36.91
✓	✓	-	-	16.63	10.45
✓	-	✓	-	57.70	45.13
✓	-	-	✓	**60.61**	**48.01**

Table 3: Ablation on the Components of SplatSSC.

Main Result

The main results on the Occ-ScanNet and Occ-ScanNet-mini benchmarks are summarized in Table 1. Our SplatSSC achieves SOTA performance, demonstrating strong robustness and fine-grained scene understanding on both benchmarks. For Occ-ScanNet, SplatSSC achieves 62.83% IoU and 51.83% mIoU, surpassing the previous SOTA RoboOcc (Zhang et al. 2025) by a substantial margin of 6.35% and 4.16%, respectively. The per-class analysis further highlights the consistent improvements brought by SplatSSC across diverse categories, from large structural elements (e.g., *walls* and *floor*) to fine-grained objects (e.g., *sofas* and *chairs*). These results underscore the strength of our synergistic design. The depth-guided initialization facilitates accurate geometric reconstruction, while our DGA ensures sharp semantic boundaries. As illustrated in the qualitative examples in Figure 5, SplatSSC yields superior 3D scene perception capabilities that surpass the previous paradigm.

$\mathcal{L}_{\text{focal}}$	$\mathcal{L}_{\text{lovasz}}$	$\mathcal{L}_{\text{scal}}^{\text{prob}}$	$\mathcal{L}_{\text{scal}}^{\text{geo}}$	$\mathcal{L}_{\text{scal}}^{\text{sem}}$	$\mathcal{L}_{\text{d}}$	IoU	mIoU
✓	✓	-	✓	✓	-	57.55	46.13
✓	✓	✓	-	-	✓	60.34	46.67
✓	✓	✓	-	✓	-	59.19	**48.28**
✓	✓	✓	-	-	-	**60.61**	48.01

Table 4: Ablation on Training Objective.

GMF	DAv2	FT-DAv2	$\delta_1 \uparrow$	RMSE ↓	C-l_1 ↓
-	✓	-	0.075	50.314	1.996
✓	✓	-	0.981	4.944	0.182
-	-	✓	0.984	3.891	0.164
✓	-	✓	**0.993**	**2.977**	**0.112**

Table 5: Ablation on Depth Branch.

Ablation Studies

Ablation studies are conducted on the Occ-ScanNet-mini dataset to assess the impact of design choice in our model.

Ablation on Gaussian Parameters. We analyze the impact of primitive count and scale range in Table 2, revealing a clear trade-off between performance and efficiency. Our setting achieves the highest semantic accuracy of 48.87% mIoU with a remarkably compact configuration of just 1200 primitives. Increasing the count to 4800 and 19200 yields marginal gains in geometric completeness but incurs higher

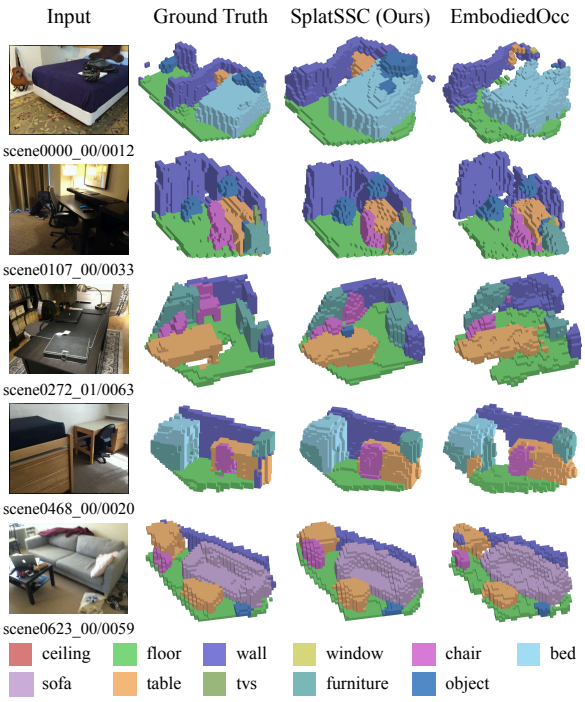

Figure 5: Qualitative results on the Occ-ScanNet-mini dataset. Our method achieves superior performance in scene completion and target object recall compared to others.

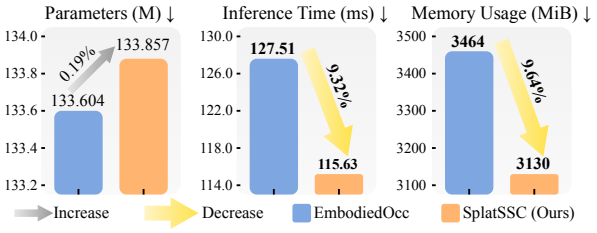

Figure 6: Efficiency Analysis.

latency and lower mIoU. The choice of scale range is equally critical. Excessively large ranges degrade accuracy and trigger Out-of-Memory (OOM) failures under dense configurations, likely due to overlaps among oversized primitives. In contrast, a moderate range [0.01, 0.16] offers the best trade-off, effectively capturing both global layouts and fine-grained details with minimal redundancy.

Ablation on Network Components. We evaluate the impact of our key components, GMF and DGA, in Table 3. The analysis first highlights the necessity of a tailored aggregation mechanism. The standard GF.agg (Huang et al. 2024) nearly fails in our sparse setting, yielding a prohibitively low 10.45% mIoU. While the more advanced GF2.agg (Huang et al. 2025) performs significantly better, our DGA still surpasses it by over 2.8% in both IoU and mIoU. This confirms that "floaters" are the key bottleneck in sparse splatting, and DGA is crucial for efficient and robust aggregation. The proposed GMF is equally important, as replacing it with a naive depth-aware baseline (Wu et al. 2025) built on Depth-Anything-V2 causes a substantial drop by more than 11% in both geometries and semantics, even when paired with our DGA. The degradation becomes more severe with other aggregators, leading to a near-collapse in performance. This demonstrates the necessity of structured geometric priors for generating informative primitives.

Ablation on Training Objective. Our validation on the training objective design is shown in Table 4. The results first confirm that the popular combination of geometry and semantic scale losses ($\mathcal{L}_{\text{scal}}^{\text{geo}}$, $\mathcal{L}_{\text{scal}}^{\text{sem}}$) is suboptimal for our framework, yielding the lowest 46.13% mIoU. The explicit depth loss $\mathcal{L}_d$ is also detrimental, as its inclusion consistently degrades both geometric and semantic scores. Furthermore, while adding semantic scale loss provides a marginal mIoU boost to a peak of 48.28%, it incurs over 1.42% IoU drop. These findings lead to our final design: a simple yet effective objective incorporating our proposed $\mathcal{L}_{\text{scal}}^{\text{prob}}$ alongside standard Focal and Lovász losses ($\mathcal{L}_{\text{focal}}$, $\mathcal{L}_{\text{lovasz}}$), which achieves the best geometric performance of 60.61% IoU while maintaining competitive semantic accuracy.

Ablation on Depth Branch. The contribution of our GMF module is validated in Table 5. The results highlight a dramatic impact of GMF on refining the geometric prior. When applied to a frozen Depth-Anything-V2 (DAv2) backbone, GMF boosts the δ_1 score by a remarkable 0.906. Furthermore, it demonstrates its capability to enhance a fine-tuned Depth-Anything-V2 (FT-DAv2) (Wu et al. 2025), pushing the δ_1 score to a new best of 0.993. This confirms that our GMF is a powerful and versatile feature refiner, essential for generating high-quality geometric representations.

Efficiency Analysis. Beyond accuracy, we evaluate the computational efficiency of SplatSSC against EmbodiedOcc, with results detailed in Figure 6. Our method demonstrates superior efficiency despite a negligible 0.19% increase in parameter count. Specifically, SplatSSC achieves a 9.32% reduction in inference latency and a 9.64% decrease in memory usage. This advantage is primarily attributed to our sparse design, which operates on significantly fewer primitives than prior works.

Conclusion

In this paper, we introduced SplatSSC, a novel framework for monocular 3D semantic scene completion. Our method addresses the critical limitations of prior object-centric approaches through two core technical contributions: 1) a depth-guided initialization strategy, powered by our group-wise multi-scale fusion module, which generates a compact and high-quality set of initial Gaussian primitives; and 2) a decoupled Gaussian aggregator that robustly resolves aggregation artifacts such as "floaters" from outlier primitives. Extensive experiments demonstrate that SplatSSC establishes a new SOTA on the Occ-ScanNet benchmark, achieving superior accuracy while simultaneously reducing latency and memory consumption.

Despite its outstanding performance, we acknowledge several limitations that offer avenues for future work as discussed in the supplementary material from our code.

Acknowledgments

This work was supported by National Research Foundation of Singapore Medium-sized Centre for Advanced Robotics Technology Innovation and Ministry of Education, Singapore, under AcRF TIER 1 Grant RG64/23.

References

Cao, A.; and de Charette, R. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3981–3991. CVPR.

Fusic, S. J.; and Sitharthan, R. 2024. Improved RRT* Algorithm-Based Path Planning for Unmanned Aerial Vehicle in a 3D Metropolitan Environment. *Unmanned Systems*, 12(05): 859–875.

Hou, J.; Li, X.; Guan, W.; Zhang, G.; Feng, D.; Du, Y.; Xue, X.; and Pu, J. 2024. FastOcc: Accelerating 3D Occupancy Prediction by Fusing the 2D Bird's-Eye View and Perspective View. In *IEEE International Conference on Robotics and Automation*, 16425–16431. ICRA.

Huang, Y.; Thammatadatrakoon, A.; Zheng, W.; Zhang, Y.; Du, D.; and Lu, J. 2025. GaussianFormer-2: Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27477–27486. CVPR.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 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. CVPR.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2024. GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction. In *Proceedings of the European Conference on Computer Vision*, volume 15085, 376–393. ECCV.

Jia, X.; Jian, S.; Tan, Y.; Che, Y.; Chen, W.; and Liang, Z. 2025. Gated Cross-Attention Network for Depth Completion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. ICASSP.

Jiang, H.; Cheng, T.; Gao, N.; Zhang, H.; Lin, T.; Liu, W.; and Wang, X. 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. CVPR.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139:1–139:14.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *14th International Conference on 3D Vision*, 239–248. 3DV.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020. Anisotropic Convolutional Networks for 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3348–3356. ICCV.

Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; and Reid, I. D. 2019. RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7693–7702. CVPR.

Li, R.; Lyu, J.; Wang, A.; Yu, R.; Wu, D.; and Xin, B. 2024. FLAGDroneRacing: An Autonomous Drone Racing System. *Unmanned Systems*, 12(06): 985–1000.

Li, Y.; Yu, Z.; Choy, C. B.; Xiao, C.; Álvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 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. CVPR.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 936–944. CVPR.

Ma, B.; Zhang, J.; Xia, Y.; and Tao, D. 2020. Auto learning attention. In *Advances in neural information processing systems*, volume 33, 1488–1500. NeurIPS.

Mei, J.; Yang, Y.; Wang, M.; Zhu, J.; Ra, J.; Ma, Y.; Li, L.; and Liu, Y. 2024. Camera-Based 3D Semantic Scene Completion With Sparse Guidance Network. *IEEE Transactions on Image Processing*, 33: 5468–5481.

Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; and Zhou, S. 2023. Occdepth: A depth-aware method for 3d semantic scene completion. arxiv:2302.13540.

Roldão, L.; de Charette, R.; and Verroust-Blondet, A. 2020. LMSCNet: Lightweight Multiscale 3D Semantic Completion. In *8th International Conference on 3D Vision*, 111–119. 3DV.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, 234–241. MICCAI.

Shi, Y.; Cheng, T.; Zhang, Q.; Liu, W.; and Wang, X. 2024. Occupancy as set of points. In *Proceedings of the European Conference on Computer Vision*, volume 15119, 72–87. ECCV.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. A. 2017. Semantic Scene Completion from a Single Depth Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 190–198. CVPR.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 6105–6114. PMLR.

Tang, P.; Wang, Z.; Wang, G.; Zheng, J.; Ren, X.; Feng, B.; and Ma, C. 2024. SparseOcc: Rethinking Sparse Latent Representation for Vision-Based Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15035–15044. CVPR.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3D: A Large-Scale 3D

Occupancy Prediction Benchmark for Autonomous Driving. In *Advances in Neural Information Processing Systems*, volume 36, 64318–64330. NeurIPS.

Tong, W.; Sima, C.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; and Li, H. 2023. Scene as Occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8372–8381. ICCV.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 5998–6008. NeurIPS.

Wang, F.; Zhang, D.; Zhang, H.; Tang, J.; and Sun, Q. 2023. Semantic Scene Completion with Cleaner Self. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 867–877. CVPR.

Wang, H.; Wei, X.; Zhang, X.; Li, J.; Bai, C.; Li, Y.; Lu, M.; Zheng, W.; and Zhang, S. 2025a. EmbodiedOcc++: Boosting Embodied 3D Occupancy Prediction with Plane Regularization and Uncertainty Sampler. In *Proceedings of the 33rd ACM International Conference on Multimedia*. MM.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotný, D. 2025b. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5294–5306. CVPR.

Wang, J.; Liu, Z.; Meng, Q.; Yan, L.; Wang, K.; Yang, J.; Liu, W.; Hou, Q.; and Cheng, M.-M. 2024a. Opus: occupancy prediction using a sparse set. In *Advances in Neural Information Processing Systems*, volume 37, 119861–119885. NeurIPS.

Wang, Y.; Chen, Y.; Liao, X.; Fan, L.; and Zhang, Z. 2024b. PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17158–17168. CVPR.

Wang, Y.; Tan, D. J.; Navab, N.; and Tombari, F. 2019. ForkNet: Multi-Branch Volumetric Semantic Completion From a Single Depth Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8607–8616. ICCV.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21672–21683. ICCV.

Wu, Y.; Zheng, W.; Zuo, S.; Huang, Y.; Zhou, J.; and Lu, J. 2025. EmbodiedOcc: Embodied 3D Occupancy Prediction for Vision-based Online Scene Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ICCV.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 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*, 3101–3109. AAAI Press.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, 21875–21911. NeurIPS.

Yang, X.; Zou, H.; Kong, X.; Huang, T.; Liu, Y.; Li, W.; Wen, F.; and Zhang, H. 2021. Semantic Segmentation-assisted Scene Completion for LiDAR Point Clouds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3555–3562. IROS.

Yi, P.; Lei, J.; Hong, Y.; and Chen, J. 2025. Embodied Intelligent Game: Models and Algorithms for Autonomous Interactions Among Heterogeneous Agents. *Unmanned Systems*, 13(05): 1365–1394.

Yu, H.; Wang, Y.; Chen, Y.; and Zhang, Z. 2024a. Monocular occupancy prediction for scalable indoor scenes. In *Proceedings of the European Conference on Computer Vision*, volume 15088, 38–54. ECCV.

Yu, Z.; Shu, C.; Deng, J.; Lu, K.; Liu, Z.; Yu, J.; Yang, D.; Li, H.; and Chen, Y. 2023. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arxiv:2311.12058.

Yu, Z.; Zhang, R.; Ying, J.; Yu, J.; Hu, X.; Luo, L.; Cao, S.-Y.; and Shen, H.-l. 2024b. Context and Geometry Aware Voxel Transformer for Semantic Scene Completion. In *Advances in Neural Information Processing Systems*, volume 37, 1531–1555. NeurIPS.

Zammit, C.; and van Kampen, E.-J. 2023. Real-time 3D UAV Path Planning in Dynamic Environments with Uncertainty. *Unmanned Systems*, 11(03): 203–219.

Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; and Yang, X. 2019. Cascaded Context Pyramid for Full-Resolution 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7800–7809. ICCV.

Zhang, Y.; Zhu, Z.; and Du, D. 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. ICCV.

Zhang, Z.; Zhang, Q.; Cui, W.; Shi, S.; Guo, Y.; Han, G.; Zhao, W.; Ren, H.; Xu, R.; and Tang, J. 2025. Roboocc: Enhancing the geometric and semantic scene understanding for robots. arxiv:2504.14604.

Zhao, L.; Wei, S.; Hays, J.; and Gan, L. 2025. GaussianFormer3D: Multi-Modal Gaussian-based Semantic Occupancy Prediction with 3D Deformable Attention. arxiv:2505.10685.

Zheng, Z.; Cao, W.; Kubota, Y.; Nakano, Y.; Gao, S.; and Suzuki, T. 2025. Robust and Energy-Efficient Torque Vectoring for a Four in-Wheel Motor Electric Vehicle Based on Sliding Mode and Model Predictive Control. *Unmanned Systems*, 13(06): 1699–1712.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the nineth International Conference on Learning Representations*. ICLR.