Dynamic Focused Masking for Autoregressive Embodied Occupancy Prediction

Yuan Sun[1] **Julio Contreras**[1] **Jorge Ortiz**[1]
[1]Rutgers, The State University of New Jersey
New Brunswick, NJ, USA 08901
ys820@soe.rutgers.edu

Abstract

Visual autoregressive modeling has recently demonstrated potential in image tasks by enabling coarse-to-fine, next-level prediction. Most indoor 3D occupancy prediction methods, however, continue to rely on dense voxel grids and convolution-heavy backbones, which incur high computational costs when applying such coarse-to-fine frameworks. In contrast, cost-efficient alternatives based on Gaussian representations—particularly in the context of multi-scale autoregression—remain underexplored. To bridge this gap, we propose DFGauss, a Dynamic Focused masking framework for multi-scale 3D Gaussian representation. Unlike conventional approaches that refine voxel volumes or 2D projections, DFGauss directly operates in the 3D Gaussian parameter space, progressively refining representations across resolutions under hierarchical supervision. Each finer-scale Gaussian is conditioned on its coarser-level counterpart, forming a scale-wise autoregressive process. To further enhance efficiency, we introduce an importance-guided refinement strategy that selectively propagates informative Gaussians across scales, enabling spatially adaptive detail modeling. Experiments on 3D occupancy benchmarks demonstrate that DFGauss achieves competitive performance, highlighting the promise of autoregressive modeling for scalable 3D occupancy prediction.

1 Introduction

With the accelerating progress in embodied intelligence and the deployment of active agents across domains such as robotics and autonomous navigation, spatial understanding has become a critical capability for intelligent systems [6, 40, 12]. To navigate indoor environments effectively, embodied agents must perform various perception tasks, among which occupancy prediction [38] plays a fundamental role in enabling agents to interpret and interact with complex real-world spaces. A key challenge in occupancy prediction lies in the trade-off between resolution and completeness: Small voxel sizes can cause holes and missing details, while larger voxels lead to over-smoothed, inaccurate geometry—issues that coarse-to-fine strategies address by progressively refining spatial resolution and structural fidelity [30]. Inspired by how humans perceive visual information—from global context to local detail in a hierarchical manner [29]—recent advances in autoregressive modeling [29, 28, 23, 20, 13] have shown great promise for addressing coarse-to-fine generation in 2D vision tasks, suggesting strong potential for enhancing 3D occupancy prediction.

While some efforts have attempted to extend autoregressive modeling to 3D volumetric representations using dense voxel grids [33, 1], these approaches often incur substantial computational overhead, limiting their scalability and generalizability. In contrast, efficient Gaussian-based representations offer a promising alternative, as they are both lightweight and capable of delivering strong performance. However, the autoregressive paradigm remains largely underexplored in the context of Gaussian-based methods, leaving a gap in effectively leveraging hierarchical modeling within computationally efficient 3D spatial frameworks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

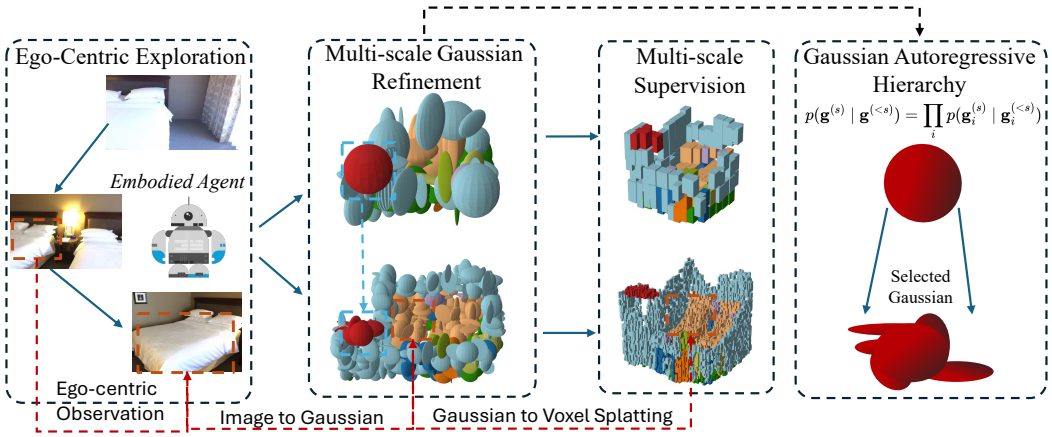

Figure 1: Overview of the proposed DFGauss framework. An embodied agent acquires egocentric RGB observations while navigating the environment, which are transformed into multi-scale Gaussian representations for autoregressive refinement. A selective masking strategy focuses computation on informative regions. In the illustrated example, the model identifies a bed (shown in peach) and ceiling (shown in dark red) through hierarchical Gaussian updates.

In this paper, we propose *DFGauss*, a Dynamic Focused masking framework for multi-scale 3D Gaussian representation (Figure 1). Our novelty lies in multi-scale Gaussian Splatting approach for 3D scene understanding that explicitly addresses the above mentioned issue through dynamic, attention-guided refinement. The core novelty of DFGauss lies in applying hierarchical autoregressive refinement directly to a sparse set of learnable 3D Gaussian primitives, avoiding reliance on dense volumetric or grid-based features. Unlike classical autoregressive models [29] that generate discrete tokens, DFGauss models spatial dependencies in a continuous space, where Gaussian parameters at each finer scale are predicted based on coarser-level outputs and corresponding encoder features. This mirrors the coarse-to-fine "next-scale prediction" strategy in Visual AutoRegressive (VAR) modeling [29], while diverging in its application to continuous Gaussian fields for structured 3D occupancy prediction. To improve efficiency, DFGauss incorporates a dynamic focused masking mechanism that selectively updates only the most informative Gaussians at each level, reducing computational cost without compromising reconstruction quality of this paradigm. Together, these components form a unified framework for efficient and expressive 3D spatial modeling, leading to the following contributions:

- A novel multi-scale autoregressive hierarchical 3D Gaussian Splatting framework that enhances 3D occupancy prediction for embodied agents.
- A coarse-to-fine supervision strategy that progressively refines Gaussian parameters across scales using multi-resolution labels.
- A dynamic focused masking mechanism that improves the efficiency of Gaussian refinement by selectively updating informative regions.
- Extensive experiments on multiple datasets demonstrating state-of-the-art performance in 3D occupancy prediction.

2 Related Work

2.1 3D Occupancy Prediction for Embodied Agents

Among various 3D perception tasks, occupancy prediction has emerged as a compact and expressive representation for modeling spatial semantics. While outdoor occupancy prediction has been extensively studied in the context of autonomous driving [33, 41, 32, 21, 9, 31, 4, 16], indoor environments remain relatively underexplored despite their critical role in embodied AI and robotics. MonoScene [3] proposes a voxel-based framework that infers occupancy from a single RGB image using a 2D-to-3D U-Net architecture with contextual priors. To enable real-time exploration, EmbodiedOcc [34] introduces a Gaussian-based memory refinement scheme. However, most existing

approaches still rely on dense 3D voxel grids [33, 19, 18], and current Gaussian-based methods remain in early stages of development. In particular, the rich features generated during Gaussian splatting have not been fully exploited. In this work, we propose a novel framework that refines sparse Gaussian features in a multi-scale autoregressive manner, combining the efficiency of Gaussian-based representations with enhanced accuracy in occupancy prediction.

2.2 Multi-scale Autoregression

Recent advances in visual autoregressive modeling (VAR) introduce a multi-scale next-resolution prediction strategy that amplifies supervision signals and enhances robustness, setting new benchmarks in 2D generation efficiency and scalability [29, 22, 11, 5]. Inspired by this, recent works have extended multi-scale modeling to 3D tasks such as occupancy prediction in autonomous driving. SurroundOcc [33], for instance, employs multi-scale 2D-3D attention to enable dense spatial reasoning and supervision. NOMAE [1] introduces a multi-scale self-supervised framework for LiDAR point clouds that focuses on localized occupancy reconstruction without modeling the full 3D volume. OctreeOcc [24] adopts an adaptive octree-based representation to support efficient and fine-grained occupancy prediction while reducing computational cost. Despite their differences, these methods rely on dense convolutional or transformer-based backbones. In contrast, the recent rise of Gaussian Splatting offers a cost-effective and compact alternative for 3D scene representation via continuous 3D Gaussian primitives. However, its integration with multi-scale optimization remains underexplored. We posit that coarse-to-fine autoregressive modeling in the scale space—where finer-scale representations are conditioned on coarser ones—provides an efficient and principled approach for structured refinement in sparse 3D representations.

2.3 Gaussian Splatting

3D Gaussian Splatting [14] has recently emerged as a compelling alternative to traditional volumetric and mesh-based rendering methods, offering real-time and high-fidelity radiance field rendering via anisotropic Gaussian primitives. Building on this foundation, a series of follow-up works aim to improve its efficiency and adaptability. Mip-Splatting [39] introduces a low-pass filtering mechanism to mitigate aliasing artifacts caused by the sampling-sensitive nature of splats. Multi-scale Gaussian Splatting [35] extends the framework to dynamic scenes by modeling object motion using MLPs. Other works focus on redundancy reduction: LightGaussian [7] and Compact3DGS [17] rank Gaussians by scale or opacity to prune uninformative primitives, significantly reducing memory and computation costs. Motivated by this line of research, we propose a dynamic focus masking mechanism that adaptively selects informative Gaussians across scales, enabling more efficient and scalable optimization in multi-scale 3D spaces, particularly for structured scene understanding, temporal consistency, and downstream prediction tasks in complex indoor environments with diverse spatial layouts and semantic variations common in embodied AI applications.

3 Method

Our model converts 2D image features into 3D Gaussian representations via cross-attention and sparse convolution. As illustrated in Figure 2, we introduce a hierarchical refinement framework that jointly optimizes multi-scale image features and Gaussian parameters across spatial resolutions. Given a monocular RGB image, a Hierarchical Feature Generation module extracts a multi-scale feature pyramid. To provide geometric cues, we incorporate a depth prediction network [36] to estimate depth maps at each level of the hierarchy. Each scale-specific feature map is then processed by a corresponding Gaussian Encoder, which predicts initial Gaussian parameters. Inspired by prior work [10, 34], we design a multi-scale Gaussian encoder that produces 3D Gaussians at different resolutions directly from the image features. These parameters are progressively refined by a Multi-Scale Gaussian Refinement module, where each level is conditioned on the output of the coarser scale via attention-based fusion, forming an autoregressive refinement process. For global Gaussian update, we adopt the same confidence-guided refinement strategy as in [34], selectively updating only the Gaussians corresponding to the original resolution. Finally, lightweight Gaussian-to-Voxel decoders apply 3D Gaussian splatting to project the refined Gaussians onto voxel grids, generating occupancy volumes at each resolution. Supervision is provided across all scales using multi-resolution 3D occupancy labels, facilitating effective learning of coarse-to-fine geometric structures. For training,

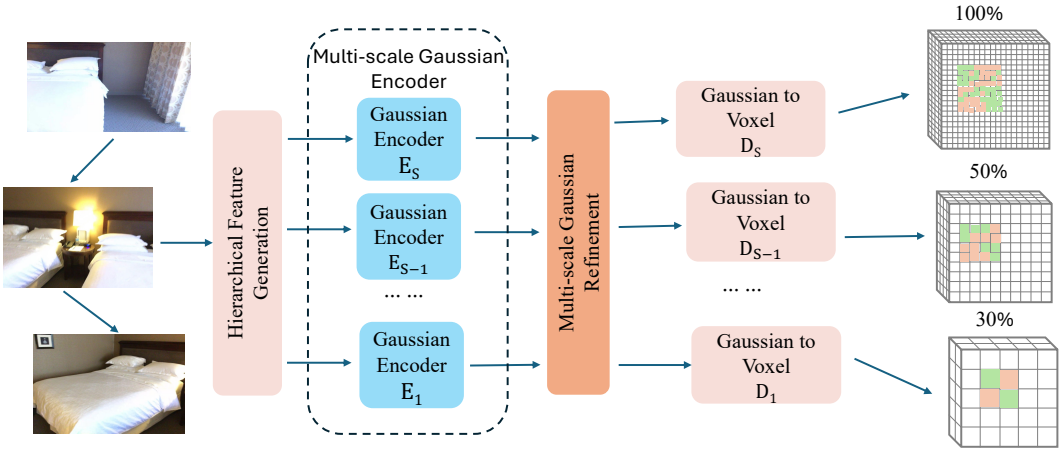

Figure 2: Overview of the DFGauss framework. Given an indoor monocular RGB image, a *Hierarchical Feature Generation* module extracts features at multiple granularity. These multi-scale features are encoded into Gaussian parameters by an *Multi-scale Gaussian Encoder*. The *Multi-Scale Gaussian Refinement* Module further optimizes these Gaussians across the hierarchy. Finally, a *Gaussian-to-Voxel* decoder predicts 3D occupancy from the multi-scale features. Supervision is provided by 3D occupancy labels at multiple resolutions.

we adopt the loss formulation from [10], combining Focal Loss, Lovász Loss, Semantic-Scale Loss, and Geometric-Scale Loss to jointly address class imbalance, boundary precision, and multi-scale consistency. In summary, our framework establishes a fully image-based Gaussian representation pipeline for monocular 3D occupancy prediction across multiple spatial granularities.

3.1 Multi-Scale Gaussian Encoder

Inspired by recent works [10, 34], we propose a Gaussian encoder that operates at each hierarchical scale of the image feature pyramid. Given image features at a specific scale s, the encoder first initializes a sparse set of 3D Gaussian primitives within camera frustum. Each Gaussian is parameterized by a tuple $\mathbf{g}_i^{(s)} = \{\mu_i, \boldsymbol{\lambda}_i, q_i, o_i, \mathbf{l}_i\}$, where $\mu_i \in \mathbb{R}^3$ is the mean position, $\boldsymbol{\lambda}_i \in \mathbb{R}^3$ are scale factors, $q_i \in \mathbb{R}^4$ is rotation quaternion, $o_i \in \mathbb{R}$ is the opacity, and $\mathbf{l}_i \in \mathbb{R}^C$ are semantic logits over C classes.

To incorporate visual cues, each Gaussian is lifted into a high-dimensional feature space via a feature alignment module, which integrates local image descriptors with geometric priors from a depth-aware branch. These features are then refined through a combination of self-attention (among Gaussians) and cross-attention (with multi-scale image features), enabling context-aware adaptation to the observed scene geometry.

To extend this to a multi-scale design, we construct a hierarchy of Gaussian encoders $\{E^{(s)}\}_{s=1}^{S}$ operating over S scales of image features $\{\mathbf{F}^{(s)}\}_{s=1}^{S}$, ordered from coarse ($s = 1$) to fine ($s = S$). Each encoder $E^{(s)}$ predicts a set of Gaussians $\mathbf{G}^{(s)}$ from its input features. Formally, we define:

$$\mathbf{G}^{(s)} = E^{(s)}(\mathbf{F}^{(s)}, D^{(s)}), \quad \mathbf{G}^{(s)} = \left\{\mathbf{g}_i^{(s)}\right\}_{i=1}^{N_s}, \quad \forall s \in \{1, \ldots, S\}, \tag{1}$$

where $D^{(s)}$ is the predicted depth map at scale s, used for geometric alignment during Gaussian initialization. Each $\mathbf{g}_i^{(s)}$ is defined in 3D space, enabling sparse yet expressive representation.

This multi-scale Gaussian encoding framework allows each level to specialize in a different spatial granularity, facilitating coarse-to-fine 3D understanding and efficient downstream refinement.

3.2 Hierarchical Multi-Scale Refinement in Gaussian Parameter Space

We propose to perform hierarchical refinement directly on the parameters of 3D Gaussians. This representation offers three key advantages: (*i*) compactness—Gaussian primitives provide a sparse,

continuous encoding of the scene without voxel quantization artifacts; (*ii*) expressiveness—each Gaussian captures both spatial location and geometric shape via learnable scale and orientation; and (*iii*) hierarchical alignment—different scales of Gaussians naturally correspond to varying levels of scene abstraction. These properties make the Gaussian parameter space a compelling domain for multi-resolution refinement, allowing the network to progressively increase spatial fidelity while preserving semantic structure. The core innovation of our approach lies in shifting the multi-scale refinement paradigm from dense voxel-based feature maps to a sparse, continuous set of learnable 3D Gaussians. By operating directly in the Gaussian parameter space, our method achieves finer geometric detail, efficient memory scaling, and hierarchical abstraction across spatial resolutions.

We explicitly formulate this process as an autoregressive refinement hierarchy for 3D scene representation, where the Gaussian set at each scale s is conditioned on all coarser levels $(1{:}s{-}1)$. Specifically, we model the hierarchical dependency as:

$$p\Big(\mathbf{G}^{(S)} \,\Big|\, \mathbf{G}^{(1:S-1)}, \hat{\mathbf{G}}^{(1:S)}\Big) = \prod_{s=1}^{S} p\Big(\mathbf{G}^{(s)} \,\Big|\, \mathbf{G}^{(1:s-1)}, \hat{\mathbf{G}}^{(s)}\Big)\,. \tag{2}$$

Here, $\hat{\mathbf{G}}^{(s)}$ denotes the initial Gaussian set predicted at scale s, and $\mathbf{G}^{(s)}$ represents its refined version obtained by updating $\hat{\mathbf{G}}^{(s)}$ with residual corrections guided by $\mathbf{G}^{(1:s-1)}$. This autoregressive structure enables structured, scale-wise propagation of geometric and semantic information.

Given the coarse-to-fine Gaussian sets $\mathbf{G}^{(1:s-1)}$ and the initial prediction $\hat{\mathbf{G}}^{(s)}$, we compute residual updates via a cross-attention fusion block:

$$\Delta\mathbf{G}^{(s)} = f_{\text{attn}}^{(s)}\Big(\mathbf{G}^{(1:s-1)}, \hat{\mathbf{G}}^{(s)}\Big)\,, \tag{3}$$

where $f_{\text{attn}}^{(s)}$ denotes a learned attention-based MLP that integrates coarse geometric priors with initial scale-s predictions for residual parameter refinement.

The refined Gaussian parameters at scale s are obtained by updating the initial predictions with learned residuals. For position, scale, opacity, and semantic logits, we have

$$\{\mu_i^{(s)}, \boldsymbol{\lambda}_i^{(s)}, o_i^{(s)}, \mathbf{l}_i^{(s)}\} = \{\hat{\mu}_i^{(s)} + \Delta\mu_i^{(s)},\ \hat{\boldsymbol{\lambda}}_i^{(s)} + \Delta\boldsymbol{\lambda}_i^{(s)},\ \hat{o}_i^{(s)} + \Delta o_i^{(s)},\ \hat{\mathbf{l}}_i^{(s)} + \Delta\mathbf{l}_i^{(s)}\}. \tag{4}$$

The rotation is updated separately by refining the initial quaternion prediction $\hat{\mathbf{q}}_i^{(s)}$ with a learned delta quaternion $\Delta\mathbf{q}_i^{(s)}$:

$$\mathbf{q}_i^{(s)} = \text{Normalize}\left(\hat{\mathbf{q}}_i^{(s)} \otimes \Delta\mathbf{q}_i^{(s)}\right), \tag{5}$$

where $\otimes$ denotes quaternion multiplication and `Normalize` enforces the unit-norm constraint. Here, $\hat{\mathbf{q}}_i^{(s)}$ is the initial rotation quaternion of the i-th Gaussian at scale s, $\Delta\mathbf{q}_i^{(s)}$ is the learned quaternion update, and $\mathbf{q}_i^{(s)}$ is the final normalized quaternion rotation after refinement.

This formulation enables structured multi-scale refinement while maintaining geometric and semantic coherence across scales and resolution levels.

3.3 Selective Gaussian Refinement Mask

To further enhance optimization efficiency, we introduce a *Selective Gaussian Refinement Mask*. The core idea is to leverage coarse-scale importance cues to identify spatially relevant regions and selectively refine only a sparse subset of fine-scale Gaussian parameters (Figure 3). By concentrating computational resources on informative areas, this mask improves both efficiency and scalability for training high-resolution 3D Gaussian fields.

Let $\mathbf{G}^{(s-1)} \in \mathbb{R}^{N_{s-1} \times D}$ and $\mathbf{G}^{(s)} \in \mathbb{R}^{N_s \times D}$ denote the Gaussian parameters at scales $s{-}1$ and s, respectively. Rather than refining fine-scale Gaussians based on all N_{s-1} coarse-scale anchors, we identify and propagate information from a smaller, semantically relevant subset.

Percentile-Based Importance Selection. Each coarse anchor $\mathbf{g}_i^{(s-1)}$ is assigned a scalar importance score $\pi_i^{(s-1)}$ via a learnable scoring function:

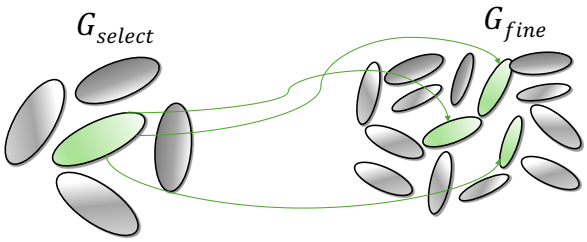

Figure 3: Illustration of the Selective Gaussian Refinement Mask. Coarse-scale Gaussians with high importance scores are first selected (left) and then used to guide the refinement of fine-scale Gaussians (right) through cross-scale attention.

$$\pi_i^{(s-1)} = \text{ScoreNet}(\mathbf{g}_i^{(s-1)}), \quad \pi_i^{(s-1)} \in \mathbb{R}. \tag{6}$$

We compute importance scores $\pi_i^{(s-1)}$ for each coarse-scale anchor using a lightweight MLP-based scoring function. We then retain the top-$\rho\%$ of anchors ranked by $\pi_i^{(s-1)}$, forming the index set:

$$\mathcal{I}_{\text{select}} = \text{Top}_\rho\left(\{\pi_i^{(s-1)}\}_{i=1}^{N_{s-1}}\right), \quad \mathbf{G}_{\text{select}}^{(s-1)} = \left\{\mathbf{g}_i^{(s-1)} \mid i \in \mathcal{I}_{\text{select}}\right\}. \tag{7}$$

Here, we apply a soft index set corresponding to the top-$\rho\%$ of Gaussians [26, 2]. Let $\mathcal{I}_{\text{select}} \in \mathbb{R}^{B\times N\times K}$ be a soft index set, where B is the batch size, N is the number of Gaussians, and K is the top-$\rho\%$ number of Gaussians. Each slice over the last dimension defines a distribution over the N elements, indicating their soft selection probability of being in the top-$\rho\%$. This formulation identifies regions requiring further refinement, as $\mathbf{G}_{\text{select}} \in \mathbb{R}^{B\times K\times D}$ is selected from the original matrix $\mathbf{G} \in \mathbb{R}^{B\times N\times D}$, where D is the feature dimension. During inference, we replace this relaxation with a discrete Top_k operator for both memory and computational efficiency.

This subset serves as a coarse-to-fine attention mask that identifies regions requiring further refinement in both geometry and semantics.

Cross-Scale Attention for Refinement. Let $\mathbf{G}_{\text{fine}}^{(s)} \in \mathbb{R}^{B\times M\times D}$ and $\mathbf{G}_{\text{select}}^{(s-1)} \in \mathbb{R}^{B\times K\times D}$ denote the fine-scale and selected coarse-scale Gaussian sets, respectively, where B is the batch size, D is the feature dimension, and M, K are the number of queried and key-value Gaussians per sample.

We apply a cross-scale attention mechanism as follows:

$$\mathbf{G}_{\text{fine}}^{(s)} = \hat{\mathbf{G}}_{\text{fine}}^{(s)} \oplus \text{softmax}\left(\frac{\hat{\mathbf{G}}_{\text{fine}}^{(s)}(\mathbf{G}_{\text{select}}^{(s-1)})^\top}{\sqrt{D}}\right)\mathbf{G}_{\text{select}}^{(s-1)}. \tag{8}$$

Here, $\oplus$ denotes the residual update operator in the Gaussian parameter space: quaternion components are composed, while position, scale, opacity, and semantic logits are updated through element-wise addition, consistent with Eqs. (4)–(5).The resulting $\mathbf{G}_{\text{fine}}^{(s)}$ retains the same shape (B, M, D).

4 Experiment

We conduct experiments on three indoor occupancy prediction benchmarks: Occ-ScanNet [38], EmbodiedOcc-ScanNet [34], and their respective smaller variants, Occ-ScanNet-mini [34] and EmbodiedOcc-ScanNet-mini [34]. Occ-ScanNet and its mini version provide monocular RGB inputs paired with voxel-level semantic labels within a frustum-aligned 3D space, supporting per-frame local occupancy prediction in a static setting. In contrast, EmbodiedOcc-ScanNet introduces a sequential and embodied formulation, where temporally continuous monocular observations enable iterative refinement of global scene understanding. Each frame is labeled with a local occupancy volume projected from a globally consistent ground-truth space, allowing both frame-wise supervision and memory-based global prediction. Following standard protocols [38, 34, 33], we report Scene

Table 1: **Local Prediction (Single-View) Results on the Occ-ScanNet dataset.**

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
MonoScene [3]	x^{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 [38]	x^{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
EmbodiedOcc [34]	x^{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
Ours	x^{rgb}	**55.28**	**42.23**	**52.95**	**43.23**	**34.20**	**43.20**	**56.73**	**63.81**	**45.66**	**35.92**	**55.23**	**44.33**	**47.03**

Table 2: **Local Prediction Results on the Occ-ScanNet-mini dataset.**

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
MonoScene [3]	x^{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 [38]	x^{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 [34]	x^{rgb}	53.80	29.10	48.70	**42.30**	38.70	42.00	62.70	60.60	48.20	33.80	**58.00**	46.50	46.40
Ours	x^{rgb}	**54.28**	**29.25**	**48.73**	38.70	**39.28**	**42.33**	**64.88**	**62.56**	**49.73**	**37.13**	57.29	**48.07**	**47.08**

Completion Intersection-over-Union (IoU) and mean Intersection-over-Union (mIoU) across semantic classes. Local metrics are computed within the camera frustum of each frame, while global metrics evaluate performance over the union of explored regions, capturing the model's ability to maintain spatial consistency over time. While our primary focus is on indoor occupancy prediction, we further assess the generalization capability of our approach on outdoor scenarios using publicly available datasets. These additional experiments underscore the broader applicability of our method. Experimental details and additional evaluation results are provided in Appendix.

4.1 Main Results

Local Prediction (Single-View) Results. In the single-view setting, DFGauss consistently outperforms prior baselines across datasets of varying scales. On the Occ-ScanNet benchmark (Table 1), our method achieves an IoU of 55.28 and mIoU of 47.03, surpassing the Gaussian-based baseline [34] by 1.33% and 1.55% respectively. This highlights the advantage of our multi-scale autoregressive Gaussian refinement. On the smaller-scale Occ-ScanNet-mini dataset (Table 2), DFGauss further improves performance with an IoU of 54.28 and mIoU of 47.08, outperforming all voxel-based and Gaussian-based baselines. These consistent gains across both full and mini variants demonstrate the robustness and scalability of our approach for local occupancy prediction. These results demonstrate that the hierarchical multi-scale design of DFGauss enables robust generalization under limited training data and across diverse model architectures and input modalities.

Global Prediction (Continuous-View) Results. In the continuous-view setting, DFGauss also outperforms the Gaussian-based baseline (Table 3). The proposed multi-scale refinement enhances scene detail modeling and facilitates better alignment between the current view and previously observed frames. On the EmbodiedOcc and EmbodiedOcc-mini datasets, DFGauss surpasses the single-scale Gaussian baseline(Table 4), achieving a 1.36% improvement in embodied mIoU and a 1.44% gain in mIoU on the mini subset. These results highlight the robustness of DFGauss and the effectiveness of hierarchical multi-scale regression in capturing fine-grained temporal and spatial context through dynamic, agent-centric environments.

4.2 Ablation Study

Component-Wise Ablation. In the component-wise ablation study (Table 5), DFGauss achieves the highest performance when both the multi-scale regression module and the dynamic masking strategy are integrated. Compared to the vanilla model, introducing the multi-scale regression alone improves

Table 3: **Global Prediction (Continuous-View) Results on the EmbodiedOcc-ScanNet Dataset.**

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
SplicingOcc	x^{rgb}	49.01	**31.60**	38.80	35.50	36.30	47.10	54.50	57.20	34.40	32.50	51.20	29.10	40.74
EmbodiedOcc[34]	x^{rgb}	51.52	22.70	44.60	**37.40**	39.00	50.10	56.70	59.70	35.40	38.40	52.00	32.90	42.53
Ours	x^{rgb}	**53.80**	26.35	**46.22**	36.73	**39.22**	**50.37**	**57.41**	**60.21**	**37.50**	**42.22**	**53.20**	**33.25**	**43.89**

Table 4: **Global Prediction Results on the EmbodiedOcc-ScanNet-mini dataset.**

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
SplicingOcc	x^{rgb}	48.80	**29.00**	37.60	37.30	26.80	44.50	66.00	52.70	40.80	36.60	54.50	27.90	41.20
EmbodiedOcc [34]	x^{rgb}	50.70	21.50	44.50	38.30	27.90	46.90	64.70	55.30	42.70	35.80	52.50	27.50	41.60
Ours	x^{rgb}	**52.32**	22.73	**44.80**	**38.70**	**30.21**	**47.12**	**65.10**	**55.62**	**43.21**	**37.83**	**55.24**	**32.83**	**43.04**

IoU by 1.85% and mIoU by 0.78%. When the dynamic mask is further applied, the gains increase to 2.71% in IoU and 1.13% in mIoU. These results indicate that the multi-scale hierarchy effectively refines the details of occupancy prediction, while the dynamic mask focuses computation on regions requiring the most refinement.

Mask Ratio Ablation. In the mask ratio ablation study (Table 6), we observe that increasing the mask ratio up to 50% consistently improves both local and global performance metrics, indicating that denser refinement improves representational quality. However, beyond this point, further increases do not yield consistent gains and may slightly degrade performance, likely due to overly sparse Gaussian selections limiting effective refinement. Additionally, lower mask ratios lead to reductions in memory usage, reflecting the efficiency benefits of selectively refining only the most informative regions.

Hierarchy Level Ablation. As shown in the hierarchy level ablation (Table 7), performance steadily improves with increasing refinement depth up to level 4, after which the gains plateau or slightly decline under a fixed masking ratio. Notably, the latency increases gradually—from 165 ms to 287 ms—as the refinement depth grows, indicating that deeper hierarchies incur only moderate computational overhead. This suggests that our Gaussian-based framework enables scalable multi-scale refinement while maintaining reasonable inference efficiency.

Table 5: Ablation on the Components

Multi-Scale	Mask	Local Metric		Global Metric	
		IoU	mIoU	IoU	mIoU
-	-	52.33	46.20	51.33	42.46
✓	-	54.32	46.65	53.04	43.21
✓	✓	**55.28**	**47.03**	**53.80**	**43.89**

Table 6: Effect of mask ratio on inference memory and performance

Ratio%	Memory	Local Metric		Global Metric	
		IoU	mIoU	IoU	mIoU
80	5328 M	54.03	46.12	**53.83**	43.47
70	5524 M	**55.82**	46.54	53.72	43.63
60	5732 M	55.28	**47.03**	53.80	**43.89**
50	5998 M	55.68	46.93	53.76	43.58
40	6179 M	54.84	46.73	53.42	43.21

4.3 Qualitative Results

Figure 4 presents qualitative results on the Occ-ScanNet dataset. Compared to the Gaussian-based baseline without autoregressive hierarchy, DFGauss captures more fine-grained structures in the target space, demonstrating the benefit of multi-scale refinement in improving spatial detail.

Table 7: Ablation on hierarchical depth; depth 1 is the baseline without refinement.

		Local Metric		Global Metric	
Depth	Latency (ms)	IoU	mIoU	IoU	mIoU
1	165	52.33	46.20	51.33	42.46
2	193	54.67	46.88	52.25	42.95
3	232	55.28	**47.03**	**53.8**	**43.89**
4	268	**55.89**	46.78	53.65	43.72
5	287	55.13	46.83	53.83	43.62

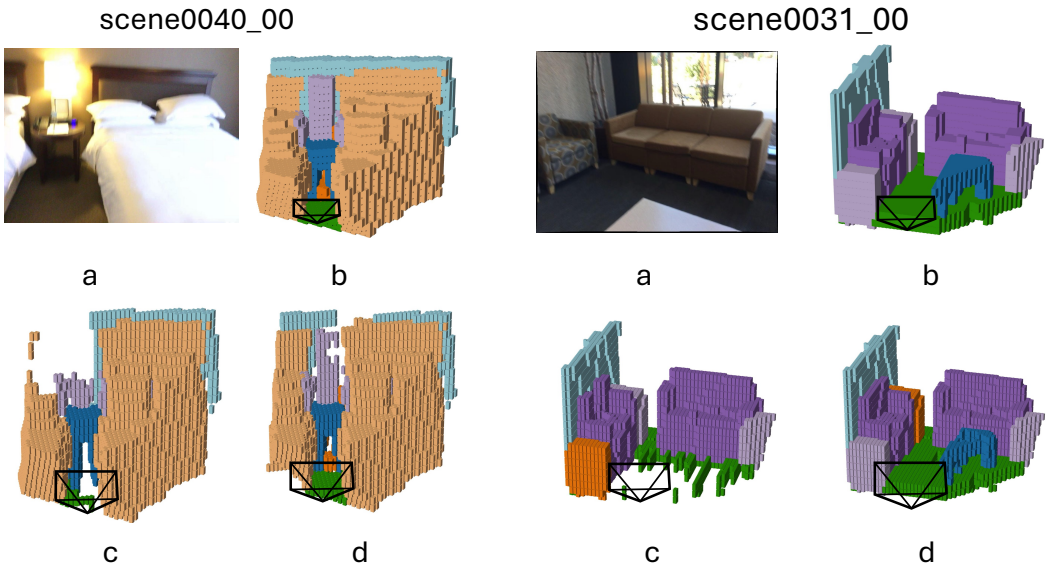

Figure 4: Qualitative results on Occ-ScanNet: (a) Input, (b) GT, (c) EmbodiedOcc, (d) DFGauss.

5 Limitation and Future Work

In this paper, we primarily focus on a Gaussian-based model architecture for monocular indoor occupancy prediction. However, extending this framework to multi-view indoor perception [34] remains an open direction and is not explored in this work. Additionally, the proposed mechanism could be applied to a broader range of point cloud tasks, such as panoptic segmentation [25, 37], semantic segmentation [27], occupancy prediction [38], and 3D scene completion [15, 8]. In future work, we plan to extend our experiments to these 3D scene understanding tasks to further investigate the potential of multi-scale Gaussian representations across diverse domains.

6 Conclusion

We present DFGauss, a multi-scale autoregressive Gaussian framework for 3D occupancy prediction. Unlike traditional approaches that rely on dense volumetric multi-scale regression, our method focuses on learning in the Gaussian parameter space across scales—a sparse and efficient representation that remains underexplored. To further enhance both accuracy and efficiency, we introduce a selective, dynamically focused refinement mask that prioritizes informative regions during hierarchical refinement. Extensive experiments demonstrate that our approach improves occupancy prediction performance while also reducing computational overhead. We hope that this framework inspires future research on efficient and expressive 3D scene understanding.

References

[1] Mohamed Abdelsamad, Michael Ulrich, Claudius Gläser, and Abhinav Valada. Multi-scale neighborhood occupancy masked autoencoder for self-supervised learning in LiDAR point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[2] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

[3] Anh-Quan Cao and Renaud de Charette. MonoScene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Junliang Chen, Huaiyuan Xu, Yi Wang, and Lap-Pui Chau. OccProphet: Pushing the efficiency frontier of camera-only 4d occupancy forecasting with an observer-forecaster-refiner framework. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[5] Liang Chen, Sinan Tan, Zefan Cai, Weichu Xie, Haozhe Zhao, Yichi Zhang, Junyang Lin, Jinze Bai, Tianyu Liu, and Baobao Chang. A spark of vision-language intelligence: 2-dimensional autoregressive transformer for efficient finegrained image generation. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[6] Shizhe Chen, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. SUGAR : Pre-training 3d visual representations for robotics. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18049–18060, 2024.

[7] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. LightGaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 140138–140158. Curran Associates, Inc., 2024.

[8] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–425, 2019.

[9] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19946–19956, 2024.

[10] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 376–393. Springer, 2024.

[11] Jiwan Hur, Dong-Jae Lee, Gyojin Han, Jaehyun Choi, Yunho Jeon, and Junmo Kim. Unlocking the capabilities of masked generative models for image synthesis via self-guidance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 130977–130999. Curran Associates, Inc., 2024.

[12] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view PointNet for 3d scene understanding. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3995–4003, 2019.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*, 2018.

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

[15] Wesley Khademi and Fuxin Li. Point-based instance completion with scene constraints. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[16] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369. Springer, 2022.

[17] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21719–21728, 2024.

[18] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers, 2024.

[19] Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, and yilun chen. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[21] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–71. Springer, 2024.

[22] Qihao Liu, Zhanpeng Zeng, Ju He, Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 133879–133907. Curran Associates, Inc., 2024.

[23] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

[24] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 79618–79641. Curran Associates, Inc., 2024.

[25] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. PanopticFusion: Online volumetric semantic mapping at the level of stuff and things. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212, 2019.

[26] Felix Petersen, Hilde Kuehne, Christian Borgelt, and Oliver Deussen. Differentiable top-k classification learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.

[28] Keyu Tian, Yi Jiang, qishuai diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[29] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 84839–84865. Curran Associates, Inc., 2024.

[30] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3d occupancy prediction benchmark for autonomous driving. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 64318–64330. Curran Associates, Inc., 2023.

[31] Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. OccGen: generative multi-modal 3d occupancy prediction for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 95–112. Springer, 2024.

[32] JiaBao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, JIE YANG, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: Occupancy prediction using a sparse set. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[33] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21729–21740, 2023.

[34] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. EmbodiedOcc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[35] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20923–20931, 2024.

[36] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 21875–21911. Curran Associates, Inc., 2024.

[37] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. GSPN: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3942–3951, 2019.

[38] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, Cham, 2024. Springer Nature Switzerland.

[39] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-Splatting: Alias-free 3d gaussian splatting. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19447–19456, 2024.

[40] Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. SAM3D: zero-shot 3d object detection via the segment anything model. *Science China Information Sciences*, 67(4), 2024.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. Our main contributions are introduced in Section 1 and elaborated in Section 3, primarily centered around a novel multi-scale autoregressive Gaussian refinement framework for 3D occupancy prediction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes,Please see Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not propose new theory in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, we provide the methodological details of our approach in Section 3, and further elaborate on the implementation specifics in Appendix. Our method is developed using publicly available datasets and can be implemented directly on top of standard baselines or off-the-shelf models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not yet published, but the supplementary material provides all details needed to reproduce our model and results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide details of the dataset splits and the hyperparameter settings in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, the baselines do not report error bars. Reproducing all baseline results and all experiments in our study with sufficient runs to report error bars is also impractical due to computational constraints.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we report these details in Appendix, including the type of compute hardware used. We also provide memory consumption and runtime statistics related information in Tables 6, 7, and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: To the best of our knowledge, our work is conducted with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: To the best of our knowledge, we do not foresee societal impacts of our work.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the used existing datasets and models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed descriptions about our method.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve crowdsourcing research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subject involved

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: the core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.