Targeted Structure Completion for Sparse-View 3D Reconstruction in Autonomous Driving

Guoqing Wang[1], Pin Tang[1], Xiangxuan Ren[1], Liping Hou[2], and Chao Ma[1], ⋆

[1] MoE Key Lab of Artificial Intelligence, Institute of AI,
Shanghai Jiao Tong University, Shanghai, China
[2] Central Research Institute, Huawei, Beijing, China
{guoqing.wang,pin.tang,bunny_renxiangxuan,chaoma}@sjtu.edu.cn
houliping1@huawei.com
Project page: https://focusgs.github.io/

Abstract. Reconstructing 3D scene structures from sparse, low-overlap observations remains a fundamental challenge in autonomous driving. Recent state-of-the-art frameworks achieve promising results by incorporating voxel-based Gaussians, but incur substantial computational redundancy due to a uniform volumetric processing strategy. To bridge the gap between the efficiency of pixel-based Gaussian methods and the structural completeness of voxel-based Gaussian approaches, we propose **FocusGS**, a simple yet effective framework that shifts the paradigm from global densification to targeted structural completion. Our central insight is that structural completion should be decoupled from deterministic regions, with computation concentrated exclusively on areas exhibiting geometric ambiguity. Specifically, FocusGS addresses the localization challenge by deriving a 3D Geometric Ambiguity Manifold to accurately isolate localized areas prone to occlusion and high geometric uncertainty. To overcome the subsequent manifold completion challenge, we design a lightweight targeted structure completion module that selectively instantiates and optimizes continuous Gaussian queries strictly within this unstructured, sparse topological subspace. Extensive experiments demonstrate that FocusGS achieves a superior efficiency-quality trade-off, advancing state-of-the-art performance on driving-centric benchmarks while naturally reducing the total number of Gaussians by ∼74% and decreasing rendering time by ∼34%.

Keywords: 3D Scene Reconstruction · Geometric Ambiguity · Targeted Structure Completion

1 Introduction

Reconstructing 3D scene structures from sparse observations is a fundamental problem in computer vision and graphics [5, 8, 37, 49, 62]. Recent feed-forward frameworks [5, 8], built upon 3D Gaussian Splatting (3DGS) [21], directly predict pixel-based 3D Gaussians from sparse inputs in a single forward pass. These

⋆ Corresponding author.

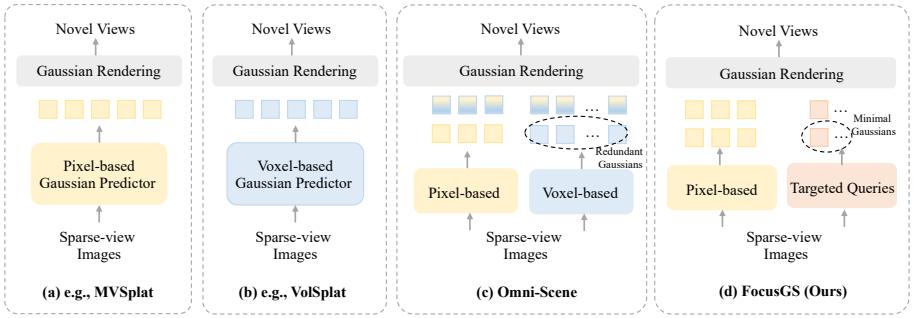

Fig. 1: Comparison of feed-forward 3D Gaussian Splatting methods. (a, b) Prior works rely on either pixel- [5,8] or voxel-based [44] predictors. (c) Omni-Scene [49] combines both but suffers from redundant Gaussians. (d) Our FocusGS introduces targeted structure completion based on pixel-based features, achieving high-fidelity reconstruction with minimal Gaussian overhead.

approaches offer remarkable efficiency and generalization, particularly in large-scale dynamic environments. However, they implicitly assume substantial overlap between input views to enforce geometric consistency. This assumption confronts the challenge in ego-centric autonomous driving, where cameras are designed for maximal coverage with minimal overlap ($< 15\%$), leading to degraded reconstruction quality.

To alleviate the reliance on cross-view overlap, voxel-based Gaussian methods [15,44,66] lift 2D image features into 3D space, thereby improving robustness to occlusion and frustum truncation (Fig. 1(b)). In particular, state-of-the-art Omni-Gaussian frameworks [49] adopt a dual-branch architecture that integrates pixel- and voxel-based Gaussians to infer and complete occluded regions explicitly. Although effective, this design introduces substantial computational overhead due to its uniform volumetric processing strategy (Fig. 1(c)), which instantiates and optimizes millions of Gaussians throughout the entire 3D space. However, in typical driving scenarios, most regions (e.g., flat roads or clear skies) are geometrically continuous and largely free of occlusion, and can therefore be adequately represented by pixel-based Gaussians alone. As a result, the volumetric branch redundantly processes extensive simple surfaces, effectively allocating a large number of voxel-based Gaussians to rectify geometric defects that are confined to only a small subset of the scene.

To quantitatively illustrate this inefficiency, we analyzed the performance of pixel-based methods across different geometric regions. As illustrated in Fig. 2, a vast majority of the environment consists of flat, deterministic areas where base representations already achieve high visual fidelity. Conversely, severe degradation occurs in a small fraction of the scene characterized by geometric ambiguity. This stark performance disparity motivates a key question: *Can we achieve the efficiency of pixel-based Gaussian methods while preserving the structural completeness as Omni-Gaussian does?* We argue that targeted structural

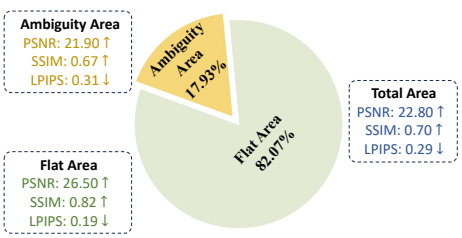

Fig. 2: Quantitative analysis of pixel-based methods across different geometric areas.

completion in critical regions provides a simple and principled solution. Rather than uniformly generating Gaussians across the entire 3D volume, the model should selectively concentrate on critical areas. We formally define these critical areas as *geometric ambiguity manifold*, i.e., a localized, sparse 3D topological subspace encompassing regions with high geometric uncertainty, such as severe depth discontinuities and occlusion boundaries. However, conducting structural completion strictly on this manifold presents two fundamental challenges. First, the localization challenge involves precisely identifying and isolating these occlusion-prone regions, which is non-trivial but essential for minimizing the number of redundant Gaussians. Second, since the ambiguity manifold is spatially irregular and sparsely distributed, traditional dense voxel grids become computationally inefficient. The targeted structure completion should therefore operate on unstructured and continuous regions without relying on dense contextual neighborhoods. Moreover, it must remain lightweight and seamlessly integrate with the base representation, avoiding the substantial overhead introduced by full voxel-based Gaussians.

To this end, we propose **FocusGS**, a simple yet effective framework for sparse-view 3D reconstruction in autonomous driving shown in Fig. 1(d). Our key insight is that geometric ambiguity in ego-centric scenes is inherently non-uniform and primarily concentrated near visibility transitions. To address the localization challenge, FocusGS first constructs pixel-based Gaussians as the base representation and subsequently derives the 3D geometric ambiguity manifold by lifting 2D ambiguity regions into a sparse 3D uncertainty subspace. This manifold serves as a precise spatial prior for incomplete geometry. To tackle the manifold completion challenge, we design a lightweight targeted structure completion module that randomly initializes a set of continuous queries strictly within this 3D uncertainty subspace. Rather than operating over the entire volume, the model selectively instantiates and optimizes targeted Gaussian queries only on the manifold. This strategy effectively decouples structural completion from deterministic regions, enabling FocusGS to resolve geometric ambiguities with a fraction of the Gaussians required by Omni-Scene.

Experiments on multiple benchmarks validate the effectiveness of FocusGS. Compared with Omni-Scene [49], FocusGS improves the efficiency-quality trade-off, reducing the total number of Gaussians by approximately **74%** and rendering time by roughly **34%**. Our main contributions are summarized as follows:

– We identify uniform volumetric processing as a critical bottleneck in existing dual-branch pipelines and leverage quantitative area analysis to motivate a novel targeted structure completion paradigm.

- We introduce FocusGS, which instantiates this paradigm by updating 3D Gaussians on the geometric ambiguity manifold by initializing targeted queries within an irregular, sparse subspace. This design allows for precise structural completion without incurring the cost of dense volumetric processing.
- Extensive experiments demonstrate that FocusGS achieves state-of-the-art reconstruction quality on nuScenes while naturally reducing redundant Gaussians by approximately 74% and rendering time by approximately 34%.

2 Related Work

2.1 3D Scene Reconstruction

Previous 3D scene reconstruction methods [2, 6, 7, 20, 25, 43, 52, 57] incorporate structural priors to predict implicit neural fields [30], light fields [38], or explicit 3D Gaussians [21] in a single forward pass. While Neural Radiance Fields (NeRF) [30] achieve high fidelity, their reliance on dense per-ray sampling imposes significant computational overhead, a bottleneck that persists even in generalizable feed-forward variants [6, 43, 57]. Although light-field approaches improve efficiency, they inherently sacrifice explicit 3D geometry [29, 36]. Consequently, explicit representations like 3D Gaussian Splatting (3DGS) [21] have emerged, utilizing differentiable rasterization to achieve real-time rendering. Building on this, recent feed-forward frameworks [5,6,19,35,54] directly predict pixel-aligned Gaussians from sparse views using geometric priors. However, these methods heavily rely on substantial cross-view overlap to enforce geometric consistency. In ego-centric autonomous driving, where cameras are designed to maximize spatial coverage, cross-view overlap is typically minimal ($< 15\%$). This causes traditional pixel-based methods to fail under severe occlusion and frustum truncation, significantly degrading reconstruction quality. Moreover, applying dense or uniform completion over the entire scene introduces substantial redundancy, as most well-observed regions already contain sufficient geometric evidence. To address these limitations, our work proposes a novel targeted structure completion paradigm to effectively resolve geometric ambiguities in sparse-view environments.

2.2 Gaussian Splatting in Autonomous Driving

The adaptation of 3D Gaussian Splatting (3DGS) [21,33,58] has driven numerous autonomous driving applications [3,9,13,14,16–18,23,24,26–28,31,32,39,41, 42,45–47,50,51,60,63,64], particularly in scene reconstruction [15,27,37,55,62]. Previous works have predominantly focused on per-scene optimization, achieving high-fidelity reconstructions by exploiting comprehensive sensor data from specific sequences. For instance, StreetGaussians [55] models dynamic urban environments by separating static backgrounds and moving vehicles into distinct Gaussian sets, employing a layered optimization strategy. Similarly, Gaussian methods for occupancy [12, 15, 67] extend scene encoding with semantic Gaussians, where each primitive serves as a flexible region of interest encapsulating

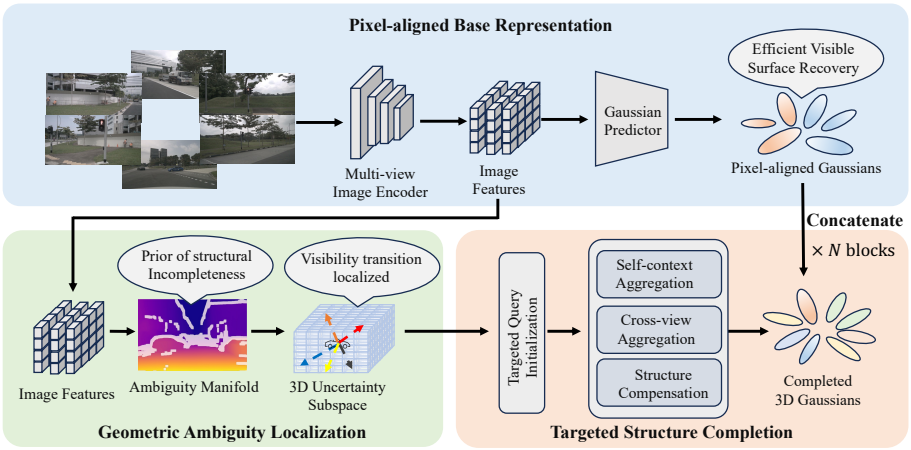

Fig. 3: The overall pipeline of FocusGS. Given sparse multi-view images, we first generate a pixel-aligned base Gaussian representation for continuous visible surfaces. To handle occlusions and depth discontinuities, we explicitly localize regions of high uncertainty by constructing a geometric ambiguity manifold. A targeted structure completion module then initializes and optimizes Gaussian queries exclusively within this derived 3D uncertainty subspace, naturally bridging the gap between pixel-based efficiency and volumetric structural completeness.

both geometric and semantic features. Recent generalizable models [5, 8] circumvent this cost by using feed-forward networks to predict Gaussians directly from sparse views. Despite these advancements, ego-centric driving reconstruction remains inherently challenged by minimal cross-view overlap and frequent occlusions. To handle this, Omni-Scene [49] introduces a dual-branch representation combining pixel- and voxel-based Gaussians. However, its uniform volumetric processing generates Gaussians across the entire 3D space indiscriminately, leading to massive computational overhead. Our FocusGS resolves this bottleneck by shifting the paradigm from global densification to targeted structural completion, efficiently targeting geometric ambiguities to preserve high-fidelity accuracy while substantially reducing the total Gaussian count.

3 Proposed Method

3.1 Overall Architecture

FocusGS introduces a **Targeted Structure Completion** paradigm to infer 3D Gaussians from sparse viewpoints as illustrated in Fig. 3. Instead of indiscriminately processing the entire 3D volume, our framework spatially decouples deterministic areas from geometrically uncertain regions, applying structural completion exclusively where it is needed. Given sparse multi-view RGB

inputs $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{N\times H\times W\times 3}$, we first extract $4\times$ down-sampled image features $\{F_i\}_{i=1}^{N} \in \mathbb{R}^{N\times \frac{H}{4}\times \frac{W}{4}\times C}$ using a 2D pretrained image backbone. We formulate the base generation as a mapping function $\mathcal{M}$ that transforms these down-sampled features into a set of pixel-aligned 3D Gaussians:

$$\mathcal{M} : \{F_i\}_{i=1}^{N} \rightarrow \{(\boldsymbol{\delta}_j,\ \boldsymbol{\alpha}_j,\ \boldsymbol{s}_j,\ \boldsymbol{q}_j,\ \boldsymbol{c}_j)\}_{j=1}^{K}, \tag{1}$$

where K denotes the total number of 3D Gaussians. The variables $\boldsymbol{\delta}_j$, $\boldsymbol{\alpha}_j$, $\boldsymbol{s}_j$, $\boldsymbol{q}_j$, and $\boldsymbol{c}_j$ represent the learned spatial offset, opacity, scale, rotation quaternion, and RGB color, respectively.

To construct this base pixel-aligned representation, the extracted image features are augmented with geometric and view-specific priors, such as Plücker ray encodings, camera embeddings, and pseudo-depth, to facilitate robust multi-view context aggregation. These geometrically enriched features are then projected to predict per-pixel depth and the corresponding 3D Gaussian attributes. To determine the precise center $\boldsymbol{\mu}_p$ of each Gaussian, we unproject the pixel from the ray origin $\mathbf{o}_p$ along the ray direction $\boldsymbol{r}_p$ using the predicted depth $\boldsymbol{d}_p$. This coarse position is subsequently refined using the learned spatial offset $\boldsymbol{\delta}_p \in \mathbb{R}^3$, formulated as:

$$\boldsymbol{\mu}_p = \boldsymbol{o}_p + \boldsymbol{d}_p\boldsymbol{r}_p + \boldsymbol{\delta}_p. \tag{2}$$

Through these steps, we obtain the base pixel-aligned Gaussians. While this base representation efficiently models deterministic regions (e.g., flat roads and clear skies), it inevitably struggles to recover complete geometry in 3D space heavily affected by occlusion and visibility transitions. To address this limitation without resorting to the computationally prohibitive uniform volumetric densification seen in prior dual-branch methods (e.g., Omni-Scene [49]), FocusGS introduces a spatially decoupled architecture. First, an ambiguity localization module extracts a 2D *geometric ambiguity manifold* from the aggregated features to explicitly identify regions requiring structural restoration. This precise spatial prior is then lifted into a sparse 3D uncertainty subspace. Finally, a lightweight targeted structural completion module selectively instantiates and optimizes additional Gaussian queries strictly within this restricted volume. This targeted design ensures high-fidelity reconstruction at complex occlusions while drastically reducing computational redundancy.

3.2 Geometric Ambiguity Localization

While pixel-based unprojection efficiently and accurately handles the vast majority of deterministic surfaces, severe reconstruction errors predominantly arise at visibility transitions, where sightlines shift abruptly between foreground and background. As demonstrated in Fig. 2, pixel-based methods achieve satisfactory performance in geometrically simple regions but suffer significant degradation in ambiguous areas. Furthermore, our empirical analysis reveals that these reconstruction errors are strongly correlated with depth discontinuities; as illustrated in Fig. 4, the spatial distribution of the error map highly overlaps with our

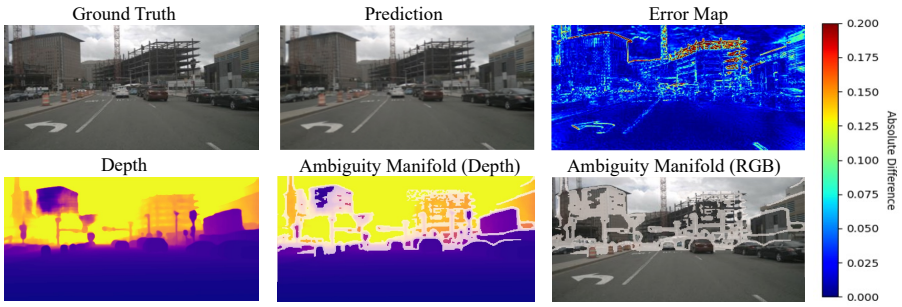

Fig. 4: Visualizing the correlation between reconstruction error and geometric ambiguity manifold. Structural artifacts mainly occur at depth boundaries, and our extracted 2D geometric ambiguity prior aligns closely with these high-error regions, effectively guiding updates to geometrically ambiguous areas.

extracted 2D prior. To address the *localization challenge*, we explicitly detect these visibility transitions to form a 2D *geometric ambiguity manifold*, which is subsequently lifted into a sparse 3D uncertainty subspace for targeted structural completion.

2D Geometric Ambiguity Manifold In surround-view images, the depth at each pixel represents the distance to the first visible surface along its camera ray. Within continuous surfaces, depth varies smoothly. However, at an occlusion boundary, the visible surface abruptly transitions from the foreground to the background, inducing a sharp depth discontinuity. We exploit the large depth gradients yielded by finite differences at these switches to effectively localize geometric ambiguity. Specifically, given the predicted depth map Z_i for each multi-view image, we compute central-difference gradients as:

$$D_x = Z_i \times K_x,\ D_y = Z_i \times K_y,\ E_i = \sqrt{D_x^2 + D_y^2}, \tag{3}$$

where $K_x = [-1, 0, 1]$, $K_y = [-1, 0, 1]^\top$, and the operator $\times$ denotes 2D convolution. We further introduce a hyperparameter τ_g to threshold the computed gradient magnitudes, isolating the strict boundaries between foreground and background:

$$\mathcal{O}_b(x, y) = \begin{cases} 1, & \text{if } E_i(x, y) \geq \tau_g \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\{x, y\}$ denote the corresponding spatial coordinates on the image plane. Because geometric uncertainty typically spans a neighborhood around the exact boundary, we apply morphological dilation with a square structuring element. This expands the strict boundaries into continuous uncertainty bands, yielding the 2D geometric ambiguity manifold $\mathcal{O}_d$ that robustly localizes occlusion-prone

or geometry-discontinuous regions. Generally, for a square structuring window of size k (with radius $r = \frac{k-1}{2}$), we obtain the dilated manifold $\mathcal{O}_d$ as:

$$\mathcal{O}_d(x, y) \;=\; \mathbf{1}\left(\sum_{i=-r}^{r} \sum_{j=-r}^{r} \mathcal{O}_b(x+i,\, y+j) \;>\; 0 \right), \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function. This dilation sets $\mathcal{O}_d(x, y)$ to 1 if at least one boundary pixel in $\mathcal{O}_b$ falls within the $k \times k$ neighborhood centered at (x, y).

3D Uncertainty Subspace Lifting Given the per-view 2D ambiguity manifolds $\mathcal{O}_d \in \{0, 1\}^{H \times W}$, we lift them into a sparse 3D neighborhood around the depth discontinuities, forming what we define as the *3D uncertainty subspace* $\mathcal{O}_{3d} \subset \mathbb{R}^3$. We project the ambiguous pixels into camera rays and subsequently transform them into the world coordinate system. For each localized pixel, we take its predicted depth d_p as the midpoint of a 3D line segment and compute the near and far endpoints, $[\mathbf{p_0}, \mathbf{p_1}]$, as:

$$\begin{aligned} \mathbf{p_0} &= \mathbf{o}_p + (d_p - \Delta_p)\mathbf{r}_p, \\ \mathbf{p_1} &= \mathbf{o}_p + (d_p + \Delta_p)\mathbf{r}_p. \end{aligned} \tag{6}$$

Here, Δ_p denotes the longitudinal thickness along the ray, which dynamically adapts to depth scale and uncertainty. It is expressed as:

$$\Delta_p = \kappa_{\text{rel}} d_p + \kappa_{\text{abs}}. \tag{7}$$

where κ_{rel} and κ_{abs} are hyperparameters controlling the allowed longitudinal variance. This formulation ensures the thickness increases linearly with depth, granting greater longitudinal uncertainty for farther points. We posit that sampling along the segment $[\mathbf{p_0}, \mathbf{p_1}]$ comprehensively covers the 3D spatial distribution of the occluded geometry. The final 3D uncertainty subspace $\mathcal{O}_{3d}$ is constructed as the union of all lifted segments across the input views:

$$\mathcal{O}_{3d} = \bigcup_{v \in \mathcal{V}} \bigcup_{\mathcal{O}_d^v(x,y)=1} \mathcal{S}_{x,y}^v, \tag{8}$$

where $S_{x,y}^v$ denotes the 3D segment originating from position $\{x, y\}$ in view v.

3.3 Targeted Structure Completion

While the base pixel-aligned representation successfully reconstructs deterministic surfaces, it inherently breaks down near visibility transitions where geometry is heavily occluded or truncated. To address this without incurring the prohibitive computational cost of uniform volumetric densification, we propose a lightweight targeted structure completion module that consists of several 3D completion blocks. Motivated by the efficiency of sparse operations [15], this module compensates for geometric incompleteness exclusively within the localized 3D uncertainty subspace $\mathcal{O}_{3d}$ derived in Section 3.2.

Uncertainty-Aware Query Sampling Instead of populating the entire 3D volume with dense voxel queries or densely sampling along every individual ray, which would lead to a highly variable computational cost depending on the scene's occlusion complexity, we maintain a strict efficiency budget by extracting a fixed total number of queries globally from the 3D uncertainty subspace $\mathcal{O}_{3d}$. Specifically, we draw exactly N_q discrete 3D points from the aggregated subspace to instantiate our localized Gaussian queries. The set of sampled query coordinates, denoted as $\mathcal{P}$, can be formulated as:

$$\mathcal{P} = \left\{ \mathbf{p}_m \in \mathcal{O}_{3d} \;\middle|\; m = 1, 2, \ldots, N_q \right\}, \quad \mathbf{p}_m \sim \mathcal{U}(\mathcal{O}_{3d}), \tag{9}$$

where N_q is a predefined hyperparameter dictating the maximum token budget, and $\mathcal{U}(\mathcal{O}_{3d})$ denotes the uniform sampling distribution over the geometry of the lifted subspace. These sampled geometric positions $\mathbf{p}_m$ serve as the initial spatial anchors for the compensatory queries, each coupled with a learnable high-dimensional feature embedding. By decoupling the query generation from the raw pixel count and confining it to a fixed budget within the geometric ambiguity manifold, we mathematically guarantee a constant and lightweight memory footprint. These N_q localized queries are subsequently optimized through N targeted completion blocks.

Self-context Aggregation To enable local geometric learning among the sampled queries, we employ 3D sparse convolution [10] for self-context aggregation. We voxelize the 3D coordinates of the sampled queries and perform 3D sparse convolutions exclusively on the occupied voxels. Since the queries are spatially restricted to the sparse $\mathcal{O}_{3d}$ subspace, this operation completely bypasses the cubic computational complexity ($\mathcal{O}(X \times Y \times Z)$) that plagues dense 3D processing pipelines. We expand the receptive field by stacking sparse convolution layers, while maintaining strict spatial sparsity to ensure maximal efficiency.

Cross-view Aggregation To resolve the geometric ambiguity inherent in occluded regions, the localized queries must gather robust multi-view context. For each 3D Gaussian query Q_{3d}, we apply deformable attention (DA) [65] onto the multi-view image feature maps. This cross-view aggregation effectively injects complementary visual cues from unoccluded viewpoints into the queries, disambiguating structures that are hidden in single ego-centric views.

Structural Compensation The goal of the structural compensation layer is to decode the final attributes of the compensatory 3D Gaussians from the enriched queries. We utilize a multi-layer perceptron (MLP) to project the updated query embeddings into explicit Gaussian properties. Specifically, we refine the 3D position (mean) of each Gaussian by predicting a residual offset relative to its initial sampled location within $\mathcal{O}_{3d}$, while the other properties (opacity, scale, rotation, and color) are directly regressed from the updated query features.

By stacking the targeted completion blocks, FocusGS generates a compact, highly localized set of supplementary Gaussians dedicated solely to structural completion. Combined with the base pixel-aligned representation, this spatially decoupled approach mitigates boundary ambiguities and enforces multi-view consistency. It achieves the structural completeness of dual-branch methods while preserving the strict efficiency of pixel-based architectures.

3.4 Training Objectives

To render a scene, we aggregate the base pixel-aligned Gaussians $\mathcal{G}_{base}$ and the targeted compensatory Gaussians $\mathcal{G}_{comp}$ to form the full scene representation, denoted as $\mathcal{G}_{final} = \mathcal{G}_{base} \cup \mathcal{G}_{comp}$. To enable appropriate supervision specifically for the targeted Gaussians, we utilize the 2D geometric ambiguity manifold $\mathcal{O}_d$ as a spatial mask. We calculate masked photometric losses as well as a masked depth loss L_{comp}^{dpt} for images and depths rendered independently from $\mathcal{G}_{comp}$. Crucially, only pixels with mask values equal to 1 are used for this targeted loss calculation, ensuring that computational resources and gradients are dedicated exclusively to ambiguity regions.

Combining this with the global photometric losses $L_{full}^{l_1}$ and L_{full}^{lpips} for novel-view images rendered from our full Gaussians $\mathcal{G}_{final}$, the overall training objective L can be derived as follows:

$$\begin{aligned} L &= L_{full}^{l_1} + \lambda_1 L_{full}^{lpips} + \lambda_2 L_{comp}, \\ L_{comp} &= L_{comp}^{l_1} + \lambda_{c_1} L_{comp}^{lpips} + \lambda_{c_2} L_{comp}^{dpt}, \end{aligned} \tag{10}$$

where λ_1 and λ_2 are weights for the LPIPS loss of $\mathcal{G}_{final}$ and the composite loss of $\mathcal{G}_{comp}$, respectively. λ_{c_1} and λ_{c_2} are weights for the masked LPIPS and depth losses of the targeted Gaussians $\mathcal{G}_{comp}$, respectively.

4 Experiments

4.1 Experimental Settings

Evaluation tasks. We follow the protocol of Omni-Scene [49] and evaluate FocusGS in two settings: the ego-centric setting on nuScenes [1] and the scene-centric setting on RealEstate10K [61]. For both datasets, we compare against the Gaussian-based methods [5, 8, 49], the light-field method AttnRend [11], and the NeRF-based method MuRF [53].

Metrics. We adopt three widely used metrics from previous 3D scene reconstruction studies [5, 8, 49] to evaluate visual quality: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [48], and learned perceptual image patch similarity (LPIPS) [59]. Higher values indicate better performance for PSNR and SSIM, whereas lower values are preferred for LPIPS. In addition, we report the Pearson correlation coefficient (PCC) to assess the geometric fidelity of reconstructed 3D scenes.

Table 1: Quantitative results of the ego-centric reconstruction task on nuScenes [1]. PCC is reported as N/A for AttnRend [11], since it does not produce an interpretable 3D structure for depth rendering.

Method	Latency(s)	PSNR↑	SSIM↑	LPIPS↓	PCC↑
AttnRend [11]	9.98	20.96	0.533	0.467	N/A
MuRF [53]	0.672	20.34	0.504	0.433	-0.332
pixelSplat [5]	0.508	21.51	0.616	0.372	0.001
MVSplat [8]	0.174	21.61	0.658	0.295	0.181
STORM [56]	-	24.56	0.752	**0.217**	0.788
SCube [34]	-	23.85	0.721	0.258	0.651
DrivingForward [40]	-	24.32	0.732	0.229	0.766
Omni-Scene [49]	0.088	24.27	0.736	0.237	0.800
FocusGS (ours)	**0.058**	**24.65**	**0.754**	0.220	**0.837**

Table 2: Quantitative results of FocusGS on RealEstate10K [61] under scene-centric reconstruction setting.

Method	PSNR↑	SSIM↑	LPIPS↓	PCC↑
AttnRend [11]	24.78	0.820	0.213	N/A
MuRF [53]	26.10	0.858	0.143	0.344
pixelSplat [5]	25.89	0.858	0.142	0.285
MVSplat [8]	26.39	0.869	0.128	0.363
Omni-Scene [49]	26.19	0.865	0.131	0.368
FocusGS	26.32	0.872	0.123	0.365
MVSplat + FocusGS	**27.02**	**0.892**	**0.118**	**0.374**

Implementation details. We implement FocusGS using the open-source Gaussian renderer [21]. Multi-view image features are extracted with a ResNet-50 backbone pre-trained using DINO [4]. For geometric ambiguity localization, depth values are clipped to the range $[0, 100]$, with the dilation kernel size and threshold as 3 and 3. We employ four targeted structure completion blocks to update within the geometric ambiguity manifold. Training is performed on two 80GB GPUs for 100k iterations with a batch size of 4 on nuScenes [1], and on a single 80GB GPU for 300k iterations with a batch size of 8 on RealEstate10K [61]. We use AdamW [22] with an initial learning rate of 1×10^{-4} and cosine decay.

4.2 Main Results

Results on nuScenes. Tab. 1 presents a comparison between FocusGS and existing baselines on nuScenes. Compared to Omni-Scene [49], specifically designed for the ego-centric setting, our approach reduces rendering latency by ∼34% while also achieving higher accuracy. Feed-forward 3DGS methods [5, 8, 11, 53] for sparse-view reconstruction perform worst, particularly on the PCC metric, as limited view overlap in ego-centric settings makes depth estimation unreliable. While Omni-Scene improves over MVSplat and pixelSplat, its voxel-based Gaussian branch has millions of Gaussians, even in well-observed regions where

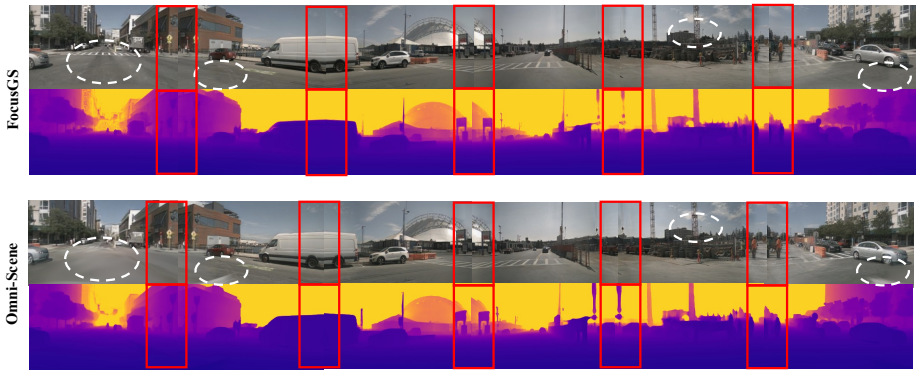

Fig. 5: Qualitative comparison between Omni-Scene [49] and FocusGS. Six rendered surround views cover the full 360° panorama with about 15% adjacent-view overlap. Red boxes mark overlap regions, while white dashed regions indicate cases where FocusGS produces sharper and more complete reconstructions.

Table 3: Ablation study of core modules under ego-centric reconstruction setting. *Random S.C.* denotes random structure completion, and *G.A. Localization* denotes geometric ambiguity localization.

	Method	PSNR↑	SSIM↑	LPIPS↓	PCC↑
(a)	Pixel-aligned Baseline	22.89	0.698	0.290	0.780
(b)	(a) + Depth Init	23.14	0.703	0.320	0.802
(c)	(b) + Random S.C.	23.40	0.708	0.306	0.810
(d)	(c) + G.A. Localization	**24.65**	**0.754**	**0.220**	**0.837**

pixel-based Gaussians suffice. In contrast, our method targets refinement only to occluded regions, substantially reducing Gaussian count while preserving performance. As highlighted by the white dashed circles in Fig. 5, Omni-Scene exhibits noticeable blurriness and structural artifacts in these specific regions, resulting in a relatively lower overall reconstruction quality. In contrast, FocusGS effectively resolves these geometric ambiguities, preserving fine details, improving geometric consistency, and achieving higher visual fidelity.

Results on RealEstate10K. To further demonstrate the effectiveness and generalization of FocusGS, we also conduct evaluations on the RealEstate10K [61] dataset, a scene-centric benchmark widely used for sparse-view reconstruction tasks. We also apply the proposed targeted structure completion to the MVSplat for further evaluation. As shown in Tab. 2, standalone FocusGS achieves competitive results, while *MVSplat+FocusGS* achieves the best performance on all metrics. We also note that feed-forward baselines, such as pixelSplat [5] and MuRF [53], although efficient, suffer from limited geometric fidelity, particularly in terms of PCC. FocusGS underscores the effectiveness of the proposed targeted structure completion strategy compared with Omni-Scene.

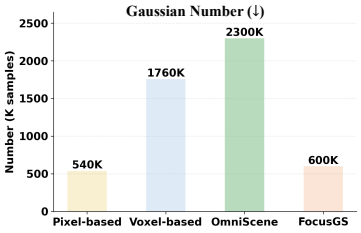

Fig. 6: Gaussian number of different methods.

Table 4: Ablation study of the blocks of targeted structure completion module under the ego-centric reconstruction setting.

Method	PSNR↑	SSIM↑	LPIPS↓	PCC↑
w/o. S.A.	24.04	0.738	0.234	0.827
w/o. C.A.	23.30	0.723	0.265	0.802
Ours	**24.65**	**0.754**	**0.220**	**0.837**

Table 5: Quantitative validation of ambiguity masks against high-error regions.

Mask Type	Recall↑	Precision↑	IoU↑	PSNR↑	SSIM↑	LPIPS↓
Error Mask (Oracle)	1.00	1.00	1.00	**27.74**	**0.812**	**0.181**
pixel-aligned (FocusGS w/o TSC)	–	–	–	22.89	0.698	0.290
Random Mask	0.43	0.16	0.13	23.42	0.703	0.285
RGB Edge Mask	0.68	0.23	0.21	23.76	0.715	0.283
Learned Mask	0.55	**0.86**	**0.50**	24.03	0.732	0.261
Ours	**0.82**	0.46	0.42	**24.65**	**0.754**	**0.220**

4.3 Ablation Study

Overall architecture. The ablations on the Depth Init, Random S.C. (Structural Completion), and G.A. (Geometric Ambiguity) Localization modules are shown in Tab. 3. *Random S.C.* brings only marginal gains over depth initialization, while G.A. Localization yields a much larger improvement, demonstrating that completion must be spatially targeted rather than randomly applied.

Targeted structure completion. We further conduct ablation studies on the components of the targeted structure completion module. We remove the self-context aggregation and cross-view aggregation layers, denoted as *w/o S.A.* and *w/o C.A.*, respectively. The structure compensation layer, which is responsible for decoding the updated features of Gaussians, cannot be ablated. As shown in Tab. 4, removing either aggregation module results in a notable performance drop. Moreover, our analysis reveals that the cross-view aggregation layer has a stronger impact on reconstruction quality compared to the self-context aggregation layer. This is because refined Gaussians iteratively obtain features both from neighboring Gaussians and from multi-view image features, and the information carried by multi-view features is substantially richer.

4.4 Further Discussion

Effectiveness of ambiguity localization. We further quantitatively evaluate different ambiguity mask designs, including random sampling, RGB-edge masks, and a learned mask. For reference, we construct an oracle error mask from ground-truth error regions, which serves as an upper bound for targeted structure completion (TSC). As shown in Tab. 5, while the learned mask achieves

Table 6: Impact of the number of Gaussians in targeted structure completion.

Nums	PSNR↑	SSIM↑	LPIPS↓	PCC↑
10k	23.79	0.747	0.231	0.830
60k	24.65	0.754	0.220	0.837
120k	24.67	0.756	0.218	0.842

Fig. 7: Failure cases of our FocusGS.

Table 7: Impact of the number of completion blocks.

Nums	PSNR↑	SSIM↑	LPIPS↓	PCC↑
1	23.76	0.747	0.224	0.828
4	**24.65**	**0.754**	**0.220**	**0.837**
6	24.56	0.744	0.226	0.841

Table 8: Impact of kernel size of 2D geometric ambiguity manifold.

Size	PSNR↑	SSIM↑	LPIPS↓	PCC↑
1	23.78	0.736	0.232	0.835
3	**24.65**	**0.754**	**0.220**	**0.837**
7	24.13	0.750	0.226	0.832

higher precision and IoU, it tends to overlook a substantial portion of high-error structural regions, limiting its effectiveness for completion. In contrast, our depth-gradient-based mask significantly improves recall, which is more critical for identifying missing or erroneous structures. Compared with the learned mask, it increases recall by +0.27 and improves PSNR by +0.62 dB, achieving a better balance between coverage and reconstruction quality.

Gaussian numbers. We conduct studies on different settings of our proposed FocusGS. When evaluating the Gaussian query number within the geometric ambiguity manifold, we observe that reconstruction quality improves as the number of queries increases, up to a critical saturation point shown in Tab. 6. Beyond this threshold, allocating additional Gaussians yields negligible geometric gains, validating that a constrained query number acts as an optimal sweet spot for maximizing visual fidelity without compromising computational efficiency.

Block numbers. Similarly, when analyzing the network depth of the completion module, we find that a moderate number of cascaded blocks is essential for sufficient context aggregation, as shown in Tab. 7. Utilizing a shallower network restricts representational capacity, whereas excessively deep architectures lead to a degradation in reconstruction quality, likely due to optimization difficulties or overfitting within the sparse feature space. Consequently, our final framework adopts these balanced configurations to ensure robust structural compensation while maintaining strict efficiency.

Kernel size. As shown in Tab. 8, we evaluate the morphological dilation kernel size defining the 2D geometric ambiguity manifold. A minimal kernel yields suboptimal quality by missing the broader uncertainty bands surrounding severe occlusions. Conversely, an excessively large kernel over-expands the manifold into flat surfaces. This over-expansion dilutes computational efficiency and degrades performance by introducing optimization interference into well-reconstructed regions. Thus, a moderate kernel size is optimal, accurately encapsulating true ambiguities to maximize visual fidelity without adding structural noise.

Failure cases. While robust in standard driving scenarios, FocusGS struggles under adverse weather conditions like rain, shown in Fig. 7. Water droplets on camera lenses introduce severe optical blur and noise directly into the input source images. This sensor-level corruption disrupts multi-view consistency, making it difficult for the network to disentangle 2D lens artifacts from the true 3D underlying geometry. Consequently, as highlighted by the red dashed circles, FocusGS erroneously bakes this input-induced ambiguity into the 3D representation, leading to blurry and distorted reconstructions in the affected regions. Addressing such weather-induced sensor noise through explicit artifact modeling remains an important direction for future work.

5 Conclusion

In this paper, we presented FocusGS, a highly efficient framework for sparse-view 3D reconstruction in ego-centric autonomous driving that resolves the computational redundancy inherent in uniform volumetric processing. To bridge the gap between pixel-based efficiency and volumetric structural completeness, we introduced a novel targeted structure completion paradigm. By formally defining a geometric ambiguity manifold to accurately localize high-uncertainty regions, FocusGS decouples structural completion from deterministic surfaces. A lightweight targeted structure completion module then optimizes continuous queries strictly within this sparse 3D topological subspace, circumventing the massive overhead of dense voxel grids. Extensive experiments demonstrate that FocusGS establishes a superior efficiency-quality trade-off. Further ablation studies validate that targeted structure completion is more effective than random completion. We hope this work can inspire future research on more adaptive and spatiotemporal Gaussian completion for complex dynamic driving scenes.

Limitations: Despite its efficiency, FocusGS still relies on the quality of the estimated geometric ambiguity manifold, which may become less reliable under highly dynamic objects, extreme lighting changes, or corrupted visual inputs. In particular, adverse weather conditions such as rain can introduce lens artifacts and severe blur, causing the model to confuse sensor-level noise with true 3D geometry; extending the current spatial completion strategy to a spatiotemporal and artifact-aware framework remains an important direction for future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62322113, 62376156), as well as the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G15).

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
2. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: CVPR. pp. 130–141 (2023)
3. Cao, A.Q., De Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR. pp. 3991–4001 (2022)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
5. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR. pp. 19457–19467 (2024)
6. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
7. Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G.: Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. IEEE TVCG (2024)
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: ECCV. pp. 370–386 (2024)
9. Chen, Z., Ye, M., Xu, S., Cao, T., Chen, Q.: Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In: ECCV. pp. 239–256. Springer (2024)
10. Contributors, S.: Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv` (2022)
11. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: CVPR. pp. 4970–4980 (2023)
12. Gan, W., Liu, F., Xu, H., Mo, N., Yokoya, N.: Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In: ICCV. pp. 28980–28990 (2025)
13. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: CVPR. pp. 17853–17862 (2023)
14. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv:2112.11790 (2021)
15. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In: ECCV. pp. 376–393. Springer (2024)
16. Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T.: Adriver-i: A general world model for autonomous driving. arXiv preprint arXiv:2311.13549 (2023)
17. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: ICCV. pp. 8340–8350 (2023)
18. Jiang, H., Liu, L., Cheng, T., Wang, X., Lin, T., Su, Z., Liu, W., Wang, X.: Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In: CVPR. pp. 11960–11970 (2025)

19. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM TOG **44**(6), 1–16 (2025)
20. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: CVPR. pp. 18365–18375 (2022)
21. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139–1 (2023)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
23. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. AAAI **37**(2), 1477–1485 (2023)
24. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV. pp. 1–18 (2022)
25. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: CVPR. pp. 7824–7833 (2022)
26. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA. pp. 2774–2781 (2023)
27. Lu, H., Xu, T., Zheng, W., Zhang, Y., Zhan, W., Du, D., Tomizuka, M., Keutzer, K., Chen, Y.: Drivingrecon: Large 4d gaussian reconstruction model for autonomous driving. arXiv preprint arXiv:2412.09043 (2024)
28. Lu, J., Huang, Z., Yang, Z., Zhang, J., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In: ECCV. pp. 329–345 (2024)
29. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM TOG **38**(4), 1–14 (2019)
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. pp. 405–421 (2020)
31. Min, C., Zhao, D., Xiao, L., Zhao, J., Xu, X., Zhu, Z., Jin, L., Li, J., Guo, Y., Xing, J., et al.: Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In: CVPR. pp. 15522–15533 (2024)
32. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. NeurIPS **37**, 56828–56858 (2024)
33. Ren, K., Jiang, L., Lu, T., Yu, M., Xu, L., Ni, Z., Dai, B.: Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. arXiv preprint arXiv:2403.17898 (2024)
34. Ren, X., Lu, Y., Liang, H., Wu, Z., Ling, H., Chen, M., Fidler, S., Williams, F., Huang, J.: Scube: Instant large-scale scene reconstruction using voxsplats. NeurIPS **37**, 97670–97698 (2024)
35. Shi, C., Shi, S., Lyu, X., Liu, C., Sheng, K., Zhang, B., Jiang, L.: Unisplat: Unified spatio-temporal fusion via 3d latent scaffolds for dynamic driving scene reconstruction. arXiv preprint arXiv:2511.04595 (2025)
36. Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. NeurIPS **34**, 19313–19325 (2021)

37. Song, Q., Li, C., Lin, H., Peng, S., Huang, R.: Adgaussian: Generalizable gaussian splatting for autonomous driving with multi-modal inputs. arXiv preprint arXiv:2504.00437 (2025)
38. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Light field neural rendering. In: CVPR. pp. 8269–8279 (2022)
39. Tang, P., Wang, Z., Wang, G., Zheng, J., Ren, X., Feng, B., Ma, C.: Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In: CVPR. pp. 15035–15044 (2024)
40. Tian, Q., Tan, X., Xie, Y., Ma, L.: Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. AAAI **39**(7), 7374–7382 (2025)
41. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In: NeurIPS. vol. 36 (2024)
42. Wang, G., Wang, Z., Tang, P., Zheng, J., Ren, X., Feng, B., Ma, C.: Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In: ECCV (2024)
43. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR. pp. 4690–4699 (2021)
44. Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Zhu, Z., Chen, D.Y., Zhuang, B.: Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297 (2025)
45. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV. pp. 55–72 (2024)
46. Wang, X., Zhu, Z., Huang, G., Wang, B., Chen, X., Lu, J.: Worlddreamer: Towards general world models for video generation via predicting masked tokens. arXiv:2401.09985 (2024)
47. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: ICCV. pp. 17850–17859 (2023)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
49. Wei, D., Li, Z., Liu, P.: Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In: CVPR. pp. 22317–22327 (2025)
50. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv:2409.03272 (2024)
51. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV. pp. 21729–21740 (2023)
52. Wu, T., Yuan, Y.J., Zhang, L.X., Yang, J., Cao, Y.P., Yan, L.Q., Gao, L.: Recent advances in 3d gaussian splatting. Computational Visual Media **10**(4), 613–642 (2024)
53. Xu, H., Chen, A., Chen, Y., Sakaridis, C., Zhang, Y., Pollefeys, M., Geiger, A., Yu, F.: Murf: multi-baseline radiance fields. In: CVPR. pp. 20041–20050 (2024)
54. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: CVPR. pp. 16453–16463 (2025)
55. Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In: ECCV. pp. 156–173 (2024)

56. Yang, J., Huang, J., Ivanovic, B., Chen, Y., Wang, Y., Li, B., You, Y., Sharma, A., Igl, M., Karkus, P., et al.: Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. In: ICLR. vol. 2025, pp. 50446–50465 (2025)
57. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR. pp. 4578–4587 (2021)
58. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: CVPR. pp. 19447–19456 (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
60. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In: ICCV. pp. 9433–9443 (2023)
61. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
62. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: CVPR. pp. 21634–21643 (2024)
63. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. arXiv preprint arXiv:2503.23463 (2025)
64. Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J.: Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. arXiv preprint arXiv:2506.13757 (2025)
65. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
66. Zhu, Z., Wang, S., Xie, J., Liu, J.j., Wang, J., Yang, J.: Voxelsplat: Dynamic gaussian splatting as an effective loss for occupancy and flow prediction. In: CVPR. pp. 6761–6771 (2025)
67. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In: CVPR. pp. 6772–6781 (2025)