

CoRoGS: Contextual Gaussian Splatting for Robust Large-Deviation View Synthesis

Xin Ma¹ Peng Lu^{1,*} Yisong Chen² Chengwei Pan³ Sheng Li²

¹Beijing University of Posts and Telecommunications

²Peking University ³Beihang University

{maxin034, lupeng}@bupt.edu.cn

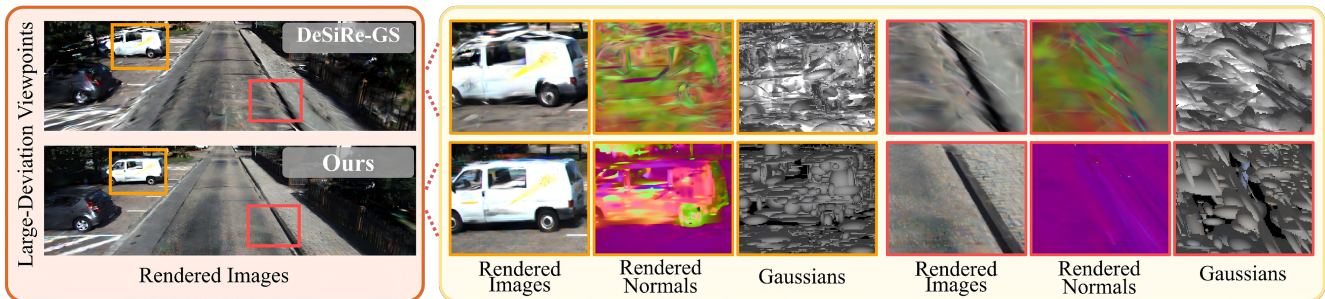


Figure 1. *Large-deviation novel view synthesis (LD-NVS) in urban scenes, evaluated on viewpoints with large deviations from the training set. The left panel presents the rendered results, while the right panel shows the corresponding zoomed-in regions, followed by the rendered normals and visualized Gaussians. Our approach shows superior rendering quality over the baseline.*

Abstract

Novel view synthesis (NVS) under large view deviations remains an underexplored challenge for 3D Gaussian Splatting (3DGS). In urban scenes with limited training coverage, models often fail to maintain geometric consistency when extrapolating to unseen viewpoints, resulting in severe distortions and degraded rendering quality. We introduce Context-Aware Gaussian Splatting (CoRoGS), a Context-aware framework for Robust large-deviation novel view synthesis (LD-NVS) that embeds contextual reasoning into 3DGS. Instead of treating Gaussians as independent primitives, CoRoGS adopts a contextual formulation that explicitly models inter-Gaussian dependencies. This representation is implemented by constructing a 3D Gaussian graph, which propagates relational geometry and semantics via message passing, resulting in context-aware Gaussian updates. To further maintain structural consistency under substantial view deviation, we incorporate a progressive graph expansion strategy that adaptively grows and prunes Gaussians, leading to more coherent and complete scene reconstructions. Extensive experiments demonstrate that CoRoGS outperforms state-of-the-art 3DGS-based methods, producing higher-quality results. We highlight that CoRoGS robustly handles a wide range of view

shifts, including lateral deviations (e.g., lane-level offsets) and cross-level transitions such as from ground-level driving views to elevated perspectives.

1. Introduction

Novel view synthesis (NVS) of urban scenes plays a vital role in autonomous driving and aerial perception systems. High-quality synthesized views not only augment multi-view training data but also enhance the robustness of downstream perception modules and scene understanding [9, 11, 16]. Recent advances in 3D Gaussian Splatting (3DGS) [8] have demonstrated impressive efficiency and photorealism for urban scene synthesis. However, achieving faithful rendering under large-deviation views (e.g., cross-lane deviation during overtaking or view transitions from ground-level to aerial view) remains an open challenge. In these scenarios, input images are captured along a single trajectory. While sparse-view 3DGS methods [2, 7] can effectively interpolate views within this trajectory, they often suffer from geometric inconsistency and appearance degradation when extrapolating to unobserved regions, resulting in visual artifacts.

Most existing 3DGS-based NVS approaches attempt to improve views generalization through geometric regularization, which can be broadly grouped into two categories:

* Corresponding author.

depth-based constraints [12, 39] that employ estimated depth maps or LiDAR priors to align Gaussian centers with underlying scene geometry; and normal-based constraints [2, 7, 21, 32] that refine local geometry by optimizing Gaussian orientations to match surface normals. While these strategies help reconstruct finer surface details, they treat Gaussians as independent primitives, lacking explicit modeling of inter-Gaussian dependencies. Consequently, geometric biases accumulate in unobserved regions, leading to overfit training views and performing poorly under large view deviations. As illustrated in Fig. 1 right, such limitations manifest as spatial deformation of Gaussians and inconsistent normal directions, revealing the absence of effective structural constraints in current formulations.

Conventional 3DGS formulations treat each Gaussian as an independent primitive, enforcing local consistency but ignoring the relational dependencies that maintain the global coherence of 3D structures. This independence limits the expressive capacity of 3DGS, particularly under large view deviations where cross-region correlations are essential for geometric and photometric consistency. To overcome this limitation, we reformulate 3D Gaussian Splatting into a contextual 3DGS framework, termed CoRoGS, which incorporates contextual reasoning as a core structural principle. Each Gaussian is refined through relational reasoning over its spatial and semantic neighborhood, allowing the model to capture both local continuity and long-range structural dependencies.

Building on this paradigm, we present a unified framework that embeds geometric and semantic context into 3DGS through Graph Neural Network. Each Gaussian serves as a node, and edges encode spatial proximity and semantic similarity, forming the foundation for graph-based contextual reasoning. The message-passing network propagates information across this graph to enable contextual Gaussian updates, achieving structural reasoning among neighboring Gaussians. Moreover, a progressive graph expansion mechanism adaptively grows Gaussians in uncovered regions and prunes redundant ones, yielding a complete and coherent scene representation. This formulation transforms 3DGS from a locally constrained representation into a context-aware reasoning system, fundamentally enhancing its robustness and generalization for large-deviation novel view synthesis. In semantically uniform regions, neighboring Gaussians reinforce smoothness and normal alignment, while in heterogeneous regions (e.g., roads, vehicles) they prevent blending across boundaries and thus maintain a globally coherent representation. This significantly improves robustness and generalization in novel view synthesis under large-deviation views.

Extensive experiments on KITTI and Waymo dataset demonstrate that the proposed method substantially enhances rendering quality, particularly for views with large

deviations. Our approach maintains geometric consistency and photometric quality over baseline methods. Furthermore, the improved robustness of CoRoGS translates to more reliable synthesized views for autonomous driving and aerial perception.

Our contributions are twofold:

- We introduce a contextual formulation of 3DGS that robustly models relational dependencies among Gaussians, establishing a novel paradigm for globally coherent and semantically consistent 3D representation.
- We present Context-Aware Gaussian Splatting, which instantiates this formulation via graph-based contextual reasoning, achieving superior rendering quality and robust generalization for large-deviation novel view synthesis.

2. Related Work

3DGS has shown impressive reconstruction quality and real-time rendering capability [4, 19, 25]. Extensions to large-scale driving scenarios [14, 15, 24] typically rely on scene partitioning and densely sampled trajectories. While effective under dense-view conditions, these methods require extensive data capture and often exhibit multi-view inconsistencies when generalized to large-deviation views.

A complementary line of research incorporates geometric priors into 3DGS to enhance reconstruction quality [20, 36]. MVS-based approaches [2, 18, 32] employ depth and normal constraints to regularize local surfaces, where LiDAR-supervised techniques [23, 26, 38] improve generalization but depend on high-fidelity sensors. Additional works leverage geometric cues such as depths [19, 37, 37, 42] or normals [21], and some adopt implicit surface reconstruction [6, 35]. However, these methods predominantly address interpolation within the observed viewpoints and are not designed for rendering views under large view deviations, which is the focus of our LD-NVS setting.

To mitigate the limitations of sparse or constrained views, several recent methods [33, 34] integrate generative priors or world models. Diffusion-based techniques [1, 7, 28] synthesize novel views followed by multi-view reconstruction, with VEGS [7] specifically targeting sparse-view scenarios relevant to LD-NVS. Despite their benefits, these methods rely on generative models with restricted camera coverage and require heavy training costs.

Graph-based extensions of 3DGS have also been explored, however, existing formulations construct graphs at the object [41] or viewpoint [40] level. SAGS [30] incorporates spatial proximity between Gaussians, but its heuristic growth strategy introduces redundant primitives and fails to capture fine-grained Gaussian dependencies, limiting efficiency and scalability. In contrast, we construct a Gaussian-level graph that models contextual dependencies for unified geometric regularization.

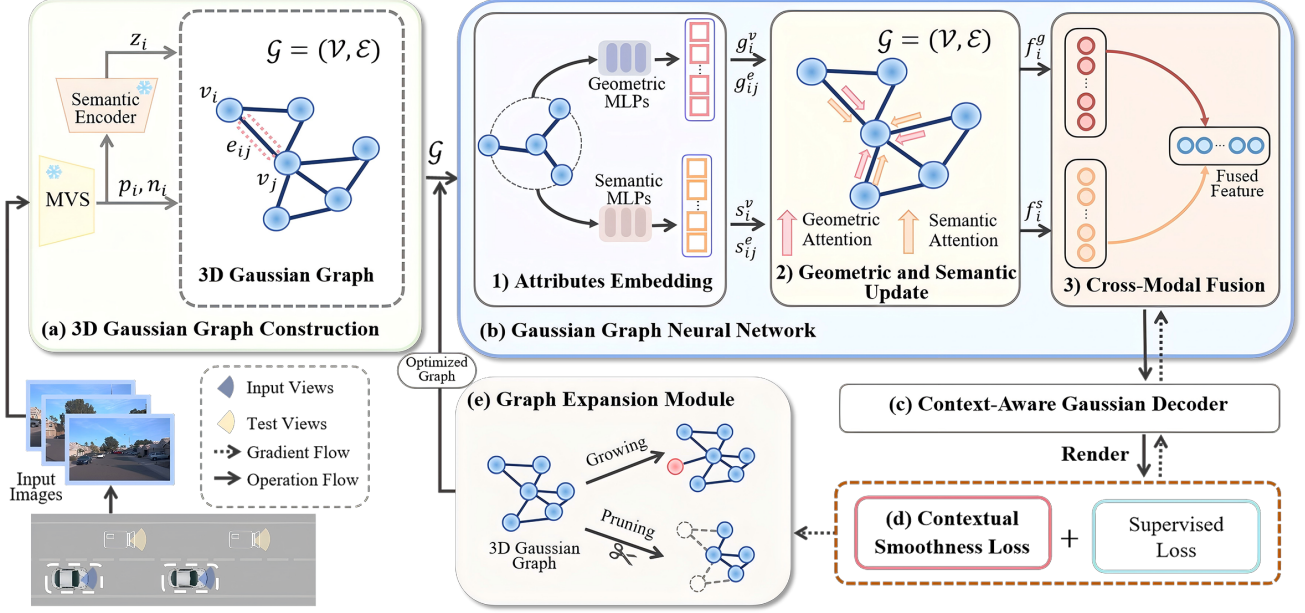


Figure 2. Overview of our CoRoGS. (a) 3D Gaussian Graph Construction module builds a graph representation of Gaussians in Sec. 3.3. (b) Gaussian Graph Neural Network in Sec. 3.4 refines each Gaussian through three submodules: 1) attributes embedding, 2) geometric and semantic update, 3) cross-modal fusion. (c) Context-aware Gaussian decoder in Sec. 3.5. (d) Contextual smoothness loss in Sec. 3.6. (e) Graph Expansion module in Sec. 3.7 iteratively refines the graph by adding nodes in uncovered regions while pruning redundant ones to alleviate artifacts caused by incomplete topology.

3. Method

3.1. Overview

We begin by introducing the Contextual 3DGS formulation, which formalizes Gaussians as a contextual representation rather than treating Gaussians as independent primitives. Building upon this formulation, we construct a 3D Gaussian Graph that integrates geometric and semantic information into a unified relational structure (Fig. 2(a)). A Gaussian Graph Neural Network then performs contextual reasoning on this graph (Fig. 2(b)), enabling each Gaussian to update its attributes by aggregating information from contextually relevant neighbors. The updated node features are decoded into Gaussian parameters through a Context-Aware Gaussian Decoder (Fig. 2(c)). This process is supervised using the proposed contextual smoothness loss (Fig. 2(d)) with other supervised losses. Finally, a Graph Expansion module iteratively refines the representation by adding Gaussians in previously uncovered regions and pruning redundant ones (Fig. 2(e)). Together, these components allow CoRoGS to model inter-Gaussian dependencies and maintain coherent scene structure, yielding high-quality renderings even under large-deviation views.

3.2. Formulation of Contextual 3DGS

Given a set of Gaussians $\mathcal{G} = \{g_i\}_{i=1}^N$, conventional 3DGS represents each Gaussian independently with parameters

$\phi_i = \{\mathbf{p}_i, \mathbf{s}_i, \mathbf{c}_i, \mathbf{q}_i, \alpha_i\}$, which collectively describe its positions, scales, colors, covariance and opacity properties. This independence assumption limits 3DGS expressive capacity, as it ignores relational dependencies among neighboring Gaussians that govern global geometric and photometric coherence.

To address this limitation, we reformulate 3DGS into a contextual representation that enables inter-Gaussian reasoning. Formally, each Gaussian is updated as a function of its own attributes and the contextual information aggregated from its neighborhood:

$$\hat{\phi}_i = \Phi(\phi_i, \mathcal{C}_i), \quad \mathcal{C}_i = \Psi(\{\phi_j \mid j \in \mathcal{N}(i)\}), \quad (1)$$

where $\mathcal{N}(i)$ denotes the neighbors of g_i , $\Psi(\cdot)$ is a contextual aggregation operator, and $\Phi(\cdot)$ fuses contextual features $\mathcal{C}_i$ into the updated representation $\hat{\phi}_i$. This formulation generalizes 3DGS by allowing Gaussian attributes to be refined through spatial and semantic correlations rather than optimized independently.

To encourage structural coherence, we introduce a contextual smoothness constraint:

$$\mathcal{L}_{\text{context}} = \sum_i \sum_{j \in \mathcal{N}(i)} w_{ij} \|\Phi(\phi_i, \mathcal{C}_i) - \Phi(\phi_j, \mathcal{C}_j)\|^2, \quad (2)$$

where w_{ij} measures the correlation strength between Gaussians i and j based on their geometric and semantic affini-

ties. This constraint enforces local consistency within homogeneous regions while allowing discontinuities at semantic boundaries, aligning local Gaussian distributions with the global scene structure. This energy function is analogous to a Markov Random Field (MRF) [13], where each node interacts only with local neighbors and global consistency emerges from local message propagation. Embedding this relational constraint into 3DGS allows the model to maintain smoothness within coherent regions while preserving sharp discontinuities at semantic boundaries, yielding a globally consistent, context-aware representation.

3.3. 3D Gaussian Graph Construction

The Gaussian graph structurally encodes the attributes of each Gaussian as well as its relationships with spatially neighboring nodes, forming the foundation for context-aware feature aggregation. We first perform Delaunay triangulation on the MVS-reconstructed point cloud [31] and take the resulting vertices and edges as Gaussian nodes and spatial connections, forming the 3D Gaussian graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v_i \in \mathcal{V} = \{v_i \mid i = 1, \dots, N\}$ corresponds to a latent representation from which Gaussians are decoded. Each node is characterized by a 3D position $\mathbf{p}_i \in \mathbb{R}^3$ and a surface normal $\mathbf{n}_i \in \mathbb{R}^3$, while $\mathcal{E} = \{e_{ij} \mid i, j = 1, \dots, N\}$ denotes adjacency relations.

Compared with k -nearest neighbor graphs, Delaunay triangulation enforces empty-circumsphere and maximal-angle criteria, yielding a more uniform and non-redundant mesh. This reduces redundant nodes and edges in regions with uneven Gaussian distribution, improving computational efficiency and reconstruction accuracy, while enhancing rendering quality.

To adapt to semantic variations and avoid artifacts at scene boundaries, each Gaussian is also assigned a semantic attribute $\mathbf{z}_i \in \mathbb{R}^{d_s}$ obtained from an encoder f_Θ [22]:

$$\mathbf{z}_i = f_\Theta(\mathbf{p}_i, \mathbf{n}_i), \quad v_i = \{\mathbf{p}_i, \mathbf{n}_i, \mathbf{z}_i\}. \quad (3)$$

Edge attributes are defined at both geometric and semantic levels. The geometric attribute of an edge is

$$\mathbf{e}_{ij}^g = [\cos(\mathbf{n}_i, \mathbf{n}_j), \|\mathbf{p}_i - \mathbf{p}_j\|_2] \in \mathbb{R}^2, \quad (4)$$

encoding the cosine similarity of normals and Euclidean distance between nodes. The semantic attribute is $\mathbf{e}_{ij}^s = \cos(\mathbf{z}_i, \mathbf{z}_j) \in \mathbb{R}$, capturing inter-node semantic consistency crucial for semantic propagation across the graph.

3.4. Gaussian Graph Neural Network

We implement the function Φ in Eq. (1) using a graph neural network (GNN) composed of three main modules: attributes embedding, geometric and semantic update, and cross-modal fusion of geometric and semantic features, as illustrated in Fig. 2(b).

3.4.1. Attributes Embedding.

Node attributes are first mapped into a high-dimensional embedding space:

$$\mathbf{g}_i^v = \mathcal{M}_v^g([\psi(\mathbf{p}_i), \psi(\mathbf{n}_i)]), \quad \mathbf{s}_i^v = \mathcal{M}_v^s(\mathbf{z}_i), \quad (5)$$

with edge embeddings derived similarly:

$$\mathbf{g}_{ij}^e = \mathcal{M}_e^g(\psi(\mathbf{e}_{ij}^g)), \quad \mathbf{s}_{ij}^e = \mathcal{M}_e^s(\psi(\mathbf{e}_{ij}^s)). \quad (6)$$

Here, $\psi(\cdot)$ denotes Fourier positional encoding, projecting node and edge attributes into a high-dimensional space to capture high-frequency information. Both $\mathcal{M}_v^{g,s}$ and $\mathcal{M}_e^{g,s}$ are implemented as two-layer MLPs for non-linear feature transformation while maintaining expressive capacity for message passing.

3.4.2. Geometric and Semantic Update

This module aggregates contextual geometric and semantic information for each node, while keeping their representations distinct. The aggregation follows a two-step procedure, where edge features are first updated and then propagated to nodes.

Edge embeddings $\mathbf{g}_{ij}^e$ and $\mathbf{s}_{ij}^e$ are updated to encode pairwise contextual relations:

$$\tilde{\mathbf{g}}_{ij}^e = \mathcal{M}_{\text{edge}}^g([\mathbf{g}_{ij}^e, \mathbf{g}_i^v, \mathbf{g}_j^v]), \quad \tilde{\mathbf{s}}_{ij}^e = \mathcal{M}_{\text{edge}}^s([\mathbf{s}_{ij}^e, \mathbf{s}_i^v, \mathbf{s}_j^v]), \quad (7)$$

where $\mathcal{M}_{\text{edge}}^{g,s}$ are learnable MLPs. This step allows each edge to capture geometric and semantic relationships between neighboring nodes.

Updated edge features are then respectively propagated to geometric and semantic nodes features via attention-based aggregation:

$$\mathbf{f}_i^g = \sum_{j \in \mathcal{N}(i)} \text{Attn}(\tilde{\mathbf{g}}_{ij}^e, \mathbf{g}_i^v), \quad \mathbf{f}_i^s = \sum_{j \in \mathcal{N}(i)} \text{Attn}(\tilde{\mathbf{s}}_{ij}^e, \mathbf{s}_i^v), \quad (8)$$

where $\text{Attn}(\cdot)$ denotes an attention-based aggregation [29].

This two-step edge-to-node update ensures that each node gathers contextually relevant geometric and semantic cues from its neighborhood, yielding a practical and expressive contextual representation.

3.4.3. Cross-Modal Fusion

Independent geometric and semantic updates capture interactions within each modal but fail to model dependencies across modalities. Semantically similar nodes may exhibit distinct geometry, and geometrically close nodes may differ in semantics. To address this, we introduce a cross-modal fusion module that refines geometric and semantic features through mutual guidance.

The independently updated features $\mathbf{f}_i^g$ and $\mathbf{f}_i^s$ are first projected into a shared latent space, $\mathbf{q}_i^g = \mathbf{W}_g \mathbf{f}_i^g$ and $\mathbf{q}_i^s = \mathbf{W}_s \mathbf{f}_i^s$, where $\mathbf{W}_g$ and $\mathbf{W}_s$ are learnable matrices,

aligning features for cross-modal interaction. Cross-modal attention is then applied, allowing each geometric feature to attend to semantically consistent neighbors and each semantic feature to attend to geometrically related ones:

$$\tilde{\mathbf{f}}_i^g = \sum_{k=1}^N \alpha_{ik}^{g \rightarrow s} \mathbf{q}_k^s, \quad \tilde{\mathbf{f}}_i^s = \sum_{k=1}^N \alpha_{ik}^{s \rightarrow g} \mathbf{q}_k^g, \quad (9)$$

with attention weights $\alpha_{ik}^{g \rightarrow s}$ and $\alpha_{ik}^{s \rightarrow g}$ computed via softmax over similarity scores, enabling mutual refinement across modals.

Finally, the refined features are fused adaptively:

$$\mathbf{z}_i^o = \eta_i \tilde{\mathbf{f}}_i^g + (1 - \eta_i) \tilde{\mathbf{f}}_i^s, \quad \eta_i = \sigma(\mathbf{W}_\eta [\tilde{\mathbf{f}}_i^g \parallel \tilde{\mathbf{f}}_i^s]), \quad (10)$$

where $\mathbf{W}_\eta$ is learnable, $\parallel$ denotes concatenation, and σ is the Sigmoid function. This fusion balances geometric detail and semantic consistency, producing a unified node representation $\mathbf{z}_i^o$.

3.5. Context-Aware Gaussian Decoder

This module generates 3D Gaussian attributes for nodes directly from fused features. Notably, each fused feature predicts attributes for N Gaussians, a design inspired by [17] and introduced here to enhance detail learning. Specifically, an MLP maps each fused feature, along with the view direction encoded via Fourier positional encoding, to the parameters of these N Gaussians: $\{(\mathbf{s}_v, \mathbf{q}_v, \alpha_v, \mathbf{c}_v)\}_{v=1}^N$ where $\mathbf{s}_v \in \mathbb{R}^3$ denotes scale, $\mathbf{q}_v \in \mathbb{R}^4$ the rotation quaternion, $\alpha_v \in \mathbb{R}$ the opacity, and $\mathbf{c}_v \in \mathbb{R}^3$ the RGB color. These attributes can be used to render images, normals and semantic maps.

3.6. Contextual Smoothness Loss

To encourage structural coherence, we implement the contextual smoothness constraint $\mathcal{L}_{\text{context}}$ in Eq. (2) as:

$$\mathcal{L}_{\text{context}} = \sum_{h=1}^{H-1} \sum_{w=1}^{W-1} \left[w_{(h,w),(h+1,w)} \|\mathcal{N}_{h+1,w} - \mathcal{N}_{h,w}\|_1 + w_{(h,w),(h,w+1)} \|\mathcal{N}_{h,w+1} - \mathcal{N}_{h,w}\|_1 \right], \quad (11)$$

where h and w denote pixel coordinates, and $\mathcal{N}$ is the rendered normal maps obtained using the method in [2]. The weight

$$w_{a,b} = \exp\left(-\frac{\|s_a - s_b\|_2^2}{2\sigma_s^2}\right) \quad (12)$$

measures the semantic similarity between pixels in the rendered images and s_i is the class probability of the i -th pixel in rendered semantic map. From Eq. (11), higher semantic similarity increases the penalty on normal differences, enforcing global geometric consistency, whereas larger semantic differences relax the constraint, improving synthesis quality near object boundaries.

Our overall training loss integrates photometric reconstruction, geometric and semantic supervision and contextual smoothness regularization defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \lambda_D \mathcal{L}_{\text{D-SSIM}} + \lambda_N \mathcal{L}_{\text{normal}} + \lambda_S \mathcal{L}_{\text{semantic}} + \lambda_C \mathcal{L}_{\text{context}} \quad (13)$$

where $\mathcal{L}_1$ denotes the photometric reconstruction loss, and $\mathcal{L}_{\text{D-SSIM}}$ follows 3DGS [8]. The normal loss $\mathcal{L}_{\text{normal}}$ encourages geometric alignment between the rendered normals and ground-truth normals extracted by [3]. The semantic loss $\mathcal{L}_{\text{semantic}}$ is defined as the cross-entropy loss between the rendered semantic images generated by the 3D Gaussian representation and semantic maps of ground-truth images obtained using LSeg [10].

3.7. Context-Aware Graph Expansion

To address the structural incompleteness introduced by MVS-based initialization, we use a Gaussian Graph Expansion mechanism that adaptively densifies the graph in regions lacking sufficient geometric support. The key idea is to monitor boundary Gaussians during training and selectively introduce new Gaussians only where the rendering process indicates missing structure.

Gaussians near graph boundaries are continuously assessed by their reconstruction sensitivity. When a boundary Gaussian exhibits persistently high optimization gradients, suggesting that its neighborhood cannot adequately account for the observed image evidence, it is split, and a new Gaussian is inserted into the uncovered region. The new Gaussian inherits the geometric and semantic attributes of its parent, ensuring continuity in local structure and appearance. To maintain spatial regularity, newly generated Gaussians are constrained by the inter-node distance bounds defined by the initial Delaunay triangulation. Candidates that violate these bounds are removed to prevent over-expansion or structural distortion.

With this gradient-driven and topology-constrained strategy, the Gaussian graph adaptively grows to fill structural gaps while preserving geometric coherence and semantic continuity. This refinement substantially improves the completeness and stability of the representation, particularly under large-deviation views.

4. Experiment and Results

4.1. Datasets and Metrics

Two large-scale urban datasets are used: KITTI [5] and Waymo Open Dataset [27]. Following previous works [2, 7], we select 6 scenes from each dataset that mainly contain static objects for evaluation. Our method is evaluated under both supervised and unsupervised settings, depending on the availability of ground-truth images. For supervised evaluation, we adopt PSNR, SSIM, and LPIPS to measure

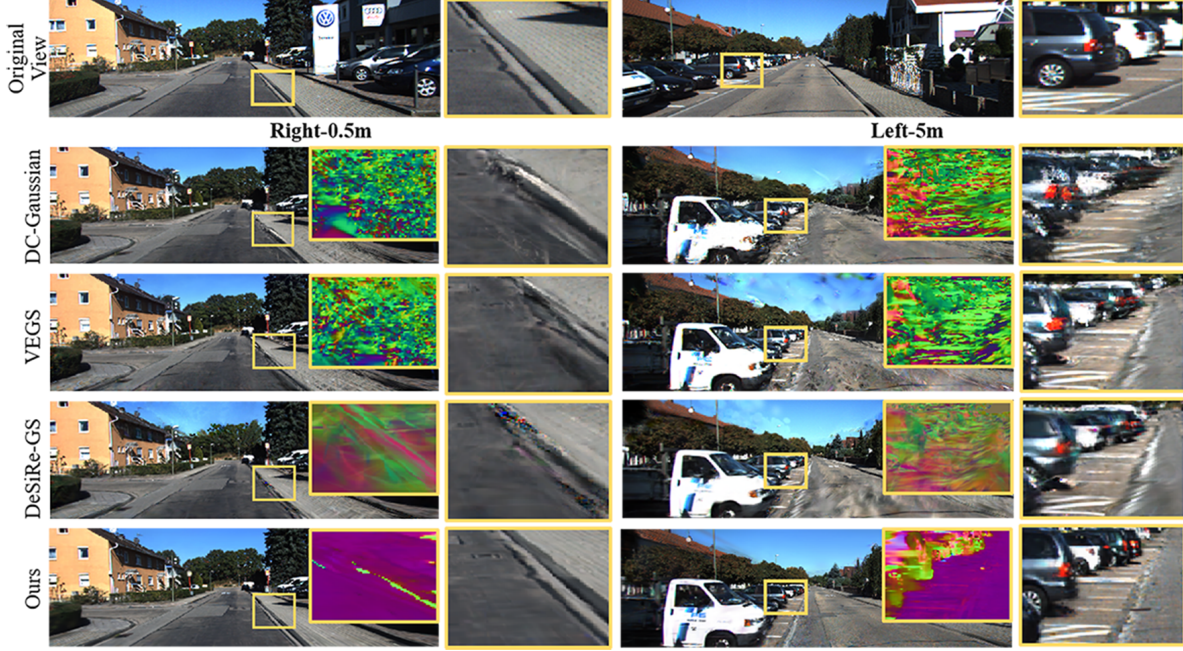


Figure 3. Qualitative comparisons on the KITTI dataset under different view deviations. Left part setting corresponds to Tab. 1, while the right corresponds to Tab. 2. Zoomed-in image patches and the associated normals demonstrate the superior rendering quality of our method compared to competing baselines.

rendering quality and perceptual similarity. For unsupervised evaluation, where ground-truth images are unavailable for pixel-level comparison, we follow [7] and employ perceptual metrics such as FID and KID to assess the generative quality and overall realism of the synthesized results. We also evaluate geometric fidelity using the Chamfer Distance (CD $\downarrow$) between rendered and original LiDAR point clouds, following the metrics established in [6].

4.2. Experimental Setup

We compare CoRoGS with several state-of-the-art 3DGS-based methods, including GaussianPro [2], SAGS [30], DC-Gaussian [32], GSDF [39], VEGS [7], and DeSiRe-GS [21]. Among these methods, GaussianPro and DC-Gaussian integrate multi-view geometric priors to reduce reconstruction errors caused by geometric inconsistencies in urban scenes. GSDF and DeSiRe-GS aim to optimize Gaussians that better align with scene surfaces, improving geometric fidelity in complex environments. In contrast, VEGS leverages diffusion-based priors to enhance novel view synthesis in regions with limited view coverage.

Supervised small-deviation setting: We use this setting with a deviation of 0.5m on the KITTI dataset, which is trained on left-camera images and evaluated on the nearly parallel right-camera views. For this setting, we evaluate performance using metrics (PSNR, SSIM, and LPIPS) to assess synthesized quality.

Unsupervised large-deviation setting: For large view deviations, these datasets lack ground truth images for

Table 1. Quantitative evaluation on the KITTI dataset using a supervised small-deviation setting with 0.5m lateral right-shift camera configuration.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CD $\downarrow$	FPS $\uparrow$	#GS(k)
GaussianPro	21.32	0.703	0.271	2.28	109	1308
SAGS	20.51	0.692	0.316	2.41	110	425
DC-GS	21.36	0.705	0.265	2.02	98	938
GSDF	21.64	0.692	0.256	2.48	102	1463
VEGS	21.85	0.699	0.275	2.31	108	1293
DeSiRe-GS	22.71	0.765	0.248	1.88	41	482
Ours	24.96	0.849	0.180	1.32	49	353

pixel-level comparison. We design six large-deviation view to evaluate model robustness with unsupervised metrics (FID and KID): (1) Left-3m and Left-5m: lateral left shift of 3m and 5m, respectively, (2) Up-1m and Up-2m: vertical shifts of 1m or 2m upward, (3) Diagonal-3m: a combined shift of 3m left and 1m upward, (4) Diagonal-5m: a combined shift of 5m left and 2m upward.

4.3. Comparison

4.3.1. Quantitative Evaluation

As shown in Table 1, we first evaluate all methods using **supervised small-deviation setting**. Our method achieves the best performance across all three metrics, with a PSNR improvement of 2.25 dB over the second-best DeSiRe-GS. In addition, our method outperforms existing baselines in

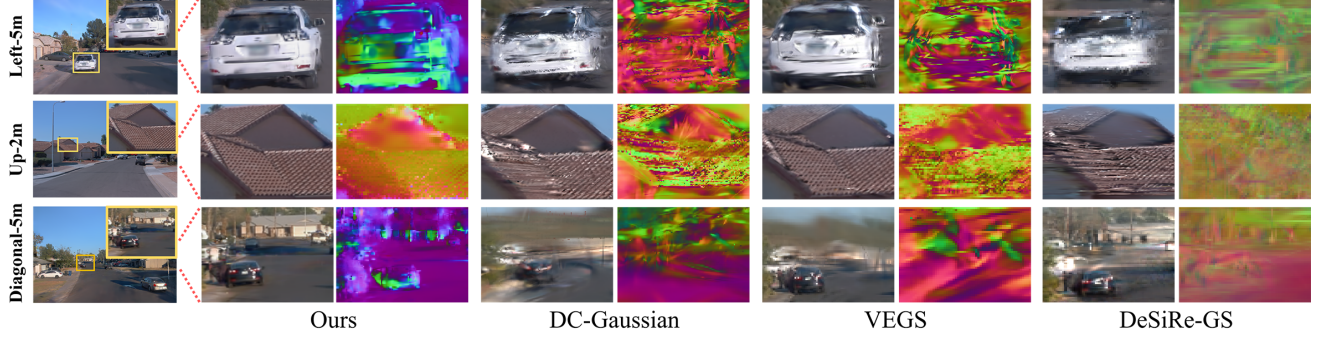


Figure 4. Qualitative comparison of rendered images and normals on the Waymo Open dataset under typical large-deviation views. Our method achieves higher visual quality than baseline approaches, consistent with the quantitative results in Tab. 2

Table 2. Quantitative results under unsupervised large-deviation settings, averaged across the KITTI and Waymo Open datasets.

Setting		Left-3m		Up-1m		Diagonal-3m	
Method	Metric	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓
GaussianPro		119.99	0.0566	99.16	0.0503	134.23	0.0791
SAGS		105.77	0.0631	107.16	0.0627	133.03	0.0807
DC-Gaussian		115.48	0.0514	94.95	0.0445	135.93	0.0786
GSDF		115.09	0.0506	96.04	0.0491	145.38	0.0829
VEGS		119.06	0.0701	108.45	0.0636	135.09	0.0801
DeSiRe-GS		104.79	0.0480	102.26	0.0479	147.87	0.0820
Ours		67.12	0.0249	72.71	0.0333	107.77	0.0413

Setting		Left-5m		Up-2m		Diagonal-5m	
Method	Metric	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓
GaussianPro		139.39	0.1034	167.49	0.1501	180.03	0.1250
SAGS		133.84	0.0819	175.94	0.1543	176.13	0.1180
DC-Gaussian		132.52	0.0911	166.42	0.1621	169.09	0.1193
GSDF		145.42	0.1133	165.36	0.1647	172.96	0.1100
VEGS		143.08	0.1089	169.04	0.1638	173.74	0.1133
DeSiRe-GS		143.28	0.0959	172.14	0.1592	175.09	0.1117
Ours		89.15	0.0378	119.19	0.0872	133.50	0.1022

terms of the CD metric, reflecting superior geometry reconstruction quality. While slower than some baselines, it requires the fewest Gaussians, achieving superior rendering quality using the minimal storage.

We further evaluate all methods using **unsupervised large-deviation setting**, with results averaged across both datasets as shown in Table 2. Detailed results for each dataset are provided in the supplementary material. For moderate shifts (Left-3m, Up-1m, and Diagonal-3m), our method achieves the best performance. Even under large shifts (Left-5m, Up-2m, and Diagonal-5m), our approach consistently outperforms DC-Gaussian [32], achieving FID improvements of 32.72%, 28.38%, and 21.04%, respectively. These results demonstrate that CoRoGS maintains

a clear and significant advantage across diverse and challenging deviation scenarios.

4.3.2. Qualitative Evaluation

We provide visual comparisons to highlight the superiority of our method. As shown in Fig. 3, under the small-deviation *Right-0.5m* setting, baselines achieve reasonable quality but miss fine details, such as the distorted curb lines marked in yellow. Our method remains clear across view shifts. Under the large-deviation *Left-5m* setting, VEGS [7] and DeSiRe-GS [21] suffer from severe artifacts on roads and vehicles, whereas our results preserve structural and texture fidelity. Consistent improvements are observed on the Waymo dataset (Fig. 4), where our method retains fine details on roofs and vehicles at varying distances, confirm-

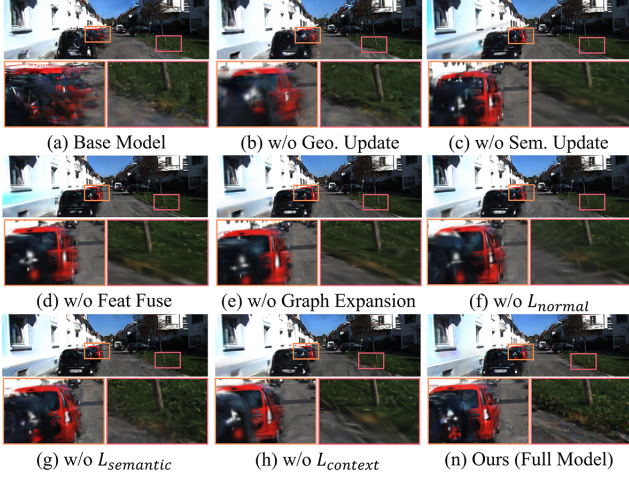


Figure 5. Ablation study on the KITTI dataset, with zoomed-in regions highlighting the visual difference.

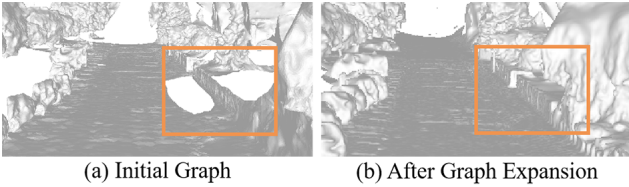


Figure 6. Visualization of initial and optimized Gaussian graph.

ing its robustness under large view deviations.

4.4. Ablations and Analysis

We conduct ablation studies under two evaluation settings on the KITTI dataset: Right-0.5m setup evaluated with PSNR, SSIM, and LPIPS metrics and Left-3m setting evaluated with FID metric.

Geometric Update Module. As comparison in Fig. 5(b)(n), removing the geometric update module in Sec. 3.4.2 introduces noticeable artifacts on vehicles and roads. Quantitative comparison in Tab. 3(b)(n) indicates the importance of aggregating information from geometrically correlated neighbors, enabling more robust modeling of complex scene geometry.

Semantic Update Module. As comparison in Fig. 5(c)(n), removing the semantic update module in Sec. 3.4.2 results in blurred renderings around vehicle and road boundaries. This degradation arises because, without semantic guidance, the model relies solely on geometric proximity, causing different semantic regions (e.g., vehicles and roads) to merge incorrectly. Incorporating semantic updates consistently improves all metrics shown in Tab. 3(c)(n).

Cross-Modal Fusion Module. Simply concatenating geometric and semantic outputs instead of using the cross-modal fusion in Sec. 3.4.3 produces visible artifacts on roads and vehicles shown in Fig. 5(c). Our fusion design

Table 3. Ablation studies on the KITTI [5] dataset.

Case Description	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
(a) Base Model	21.25	0.6889	0.2691	123.80
(b) w/o Geo. Update	23.25	0.7963	0.1985	83.62
(c) w/o Sem. Update	23.69	0.7826	0.1936	86.93
(d) w/o Fusion	23.93	0.8026	0.1965	86.34
(e) w/o Expansion	24.03	0.8136	0.1902	89.36
(f) w/o L_{normal}	23.85	0.8025	0.1928	79.68
(g) w/o $L_{semantic}$	23.79	0.7963	0.1949	82.96
(h) w/o $L_{context}$	23.32	0.7935	0.1905	89.65
(n) Ours	24.96	0.8492	0.1805	68.25

improves PSNR by 1.03 dB indicated in Tab. 3(d)(n).

Context-Aware Graph Expansion. Removing graph expansion in Sec. 3.7 leads to a consistent decline across all metrics, including a PSNR drop of 0.93 dB as shown in Tab. 3(e)(n). Visualizations in Fig. 6 show that graph expansion completes missing regions while maintaining topology. Our graph expansion enables robust surface modeling and improved rendering quality.

Loss Functions. As shown in Tab. 3(f)(g) and (h), removing each loss independently degrades rendering quality compared to our full model (case (n)). Notably, our contextual loss proposed in Sec. 3.6 provides substantial improvement, and when all three losses are combined in our full model (case (n)), the method achieves the best results.

5. Conclusion

We introduced CoRoGS, a context-aware Gaussian splatting framework that advances large-deviation novel view synthesis by moving beyond independent Gaussian optimization toward a contextual representation of 3DGS. Our formulation establishes a mechanism for modeling dependencies among Gaussians, providing a theoretical foundation for contextual Gaussian updates through a unified graph-based architecture. Building on this formulation, CoRoGS achieves photorealistic and geometrically consistent renderings under a broad range of viewpoint deviations, including large lateral offsets and substantial cross-level transitions. We highlight that jointly modeling structural context and adaptive Gaussian representation substantially improves rendering quality in real-world NVS scenarios, demonstrating strong applicability to autonomous driving and low-altitude aerial perception. Although effective, our CoRoGS still inherits limitations in handling dynamic objects and large rotations. Future work will strengthen the framework by extending contextual reasoning to dynamic Gaussian fields and incorporating rotation-aware structural constraints, further enhancing the capability and generality of context-driven 3D scene representations.

Acknowledgments

This work was supported in part by National Key R&D Program of China 2022ZD0160805.

References

- [1] Yuedong Chen, Chuanxia Zheng, Haoifei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024. 2
- [2] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8123–8140, 2024. 1, 2, 5, 6
- [3] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 5
- [4] Runze Fan, Jian Wu, Xuehuai Shi, Lizhi Zhao, Qixiang Ma, and Lili Wang. Fov-gs: Foveated 3d gaussian splatting for dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2
- [5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 5, 8
- [6] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2, 6
- [7] Sungwon Hwang, Min-Jung Kim, Taewoong Kang, Jayeon Kang, and Jaegul Choo. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *European Conference on Computer Vision*, pages 1–18, 2024. 1, 2, 5, 6, 7
- [8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1, 5
- [9] Xiaohan Lei, Min Wang, Wengang Zhou, and Houqiang Li. Gaussnav: Gaussian splatting for visual navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4108–4121, 2025. 1
- [10] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 5
- [11] Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14078–14088, 2025. 1
- [12] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. 2
- [13] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009. 4
- [14] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 2
- [15] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 2
- [16] Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Toward real-world bev perception: Depth uncertainty estimation via gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17124–17133, 2025. 1
- [17] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 5
- [18] Xin Ma, Jiguang Zhang, Peng Lu, Shibiao Xu, and Chengwei Pan. Novel view synthesis under large-deviation viewpoint for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6000–6008, 2025. 2
- [19] Sheng Miao, Jiaxin Huang, Dongfeng Bai, Xu Yan, Hongyu Zhou, Yue Wang, Bingbing Liu, Andreas Geiger, and Yiyi Liao. Evolsplat: Efficient volume-based gaussian splatting for urban view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11286–11296, 2025. 2
- [20] Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21600–21609, 2025. 2
- [21] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6782–6791, 2025. 2, 6, 7
- [22] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4
- [23] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas

- Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 2
- [24] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [25] Mingwen Shao, Yuanjian Qiao, Kai Zhang, and Lingzhuang Meng. Frequency-aware uncertainty gaussian splatting for dynamic scene reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2
- [26] Qi Song, Chenghong Li, Haotong Lin, Sida Peng, and Rui Huang. Adgaussian: Generalizable gaussian splatting for autonomous driving with multi-modal inputs. *arXiv preprint arXiv:2504.00437*, 2025. 2
- [27] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 5
- [28] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5283–5293, 2025. 2
- [29] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:4, 2018. 4
- [30] Evangelos Ververas, Rolandos Alexandros Potamias, Jifei Song, Jiankang Deng, and Stefanos Zafeiriou. Sags: Structure-aware 3d gaussian splatting. In *European Conference on Computer Vision*, pages 221–238. Springer, 2024. 2, 6
- [31] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):889–901, 2011. 4
- [32] Linhan Wang, Kai Cheng, Shuo Lei, Shengkun Wang, Wei Yin, Chenyang Lei, Xiaoxiao Long, and Chang-Tien Lu. Dc-gaussian: Improving 3d gaussian splatting for reflective dash cam videos. *Advances in Neural Information Processing Systems*, 37:99898–99920, 2024. 2, 6, 7
- [33] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difx3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26024–26035, 2025. 2
- [34] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21551–21561, 2024. 2
- [35] Haodong Xiang, Xinghui Li, Kai Cheng, Xiansong Lai, Wanting Zhang, Zhichao Liao, Long Zeng, and Xueping Liu. Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction. In *2025 IEEE International Conference on Robotics and Automation*, pages 2686–2693. IEEE, 2025. 2
- [36] Yuru Xiao, Deming Zhai, Wenbo Zhao, Kui Jiang, Junjun Jiang, and Xianming Liu. Mcgs: Multiview consistency enhancement for sparse-view 3d gaussian radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [37] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting. *eprint arXiv: 2312.00206*, 2023. 2
- [38] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2
- [39] Mulin Yu, Tao Lu, Linning Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. *Advances in Neural Information Processing Systems*, 37:129507–129530, 2024. 2, 6
- [40] Shengjun Zhang, Xin Fei, Fangfu Liu, Haixu Song, and Yueqi Duan. Gaussian graph network: Learning efficient and generalizable gaussian representations from multi-view images. *Advances in Neural Information Processing Systems*, 37:50361–50380, 2024. 2
- [41] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21634–21643, 2024. 2
- [42] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, pages 145–163. Springer, 2024. 2