

COSMOS: Coherent SuperGaussian Modeling with Spatial Priors for Sparse-View 3D Splatting

Chaeyoung Jeong¹, Kwangsu Kim^{2*}

¹Department of Artificial Intelligence, Sungkyunkwan University

²Department of Computer Science and Engineering, Sungkyunkwan University
pcc1016@skku.edu, kim.kwangsu@skku.edu

Abstract

3D Gaussian Splatting (3DGS) has recently emerged as a promising approach for 3D reconstruction, providing explicit, point-based representations and enabling high-quality real-time rendering. However, when trained with sparse input views, 3DGS suffers from overfitting and structural degradation, leading to poor generalization on novel views. This limitation arises from its optimization relying solely on photometric loss without incorporating any 3D structure priors. To address this issue, we propose Coherent supergaussian Modeling with Spatial Priors (COSMOS). Inspired by the concept of superpoints from 3D segmentation, COSMOS introduces 3D structure priors by newly defining supergaussian groupings of Gaussians based on local geometric cues and appearance features. To this end, COSMOS applies inter-group global self-attention across supergaussian groups and sparse local attention among individual Gaussians, enabling the integration of global and local spatial information. These structure-aware features are then used for predicting Gaussian attributes, facilitating more consistent 3D reconstructions. Furthermore, by leveraging supergaussian-based grouping, COSMOS enforces an intra-group positional regularization to maintain structural coherence and suppress floaters, thereby enhancing training stability under sparse-view conditions. Our experiments on Blender and DTU show that COSMOS surpasses state-of-the-art methods in sparse-view settings without any external depth supervision.

1 Introduction

3D scene reconstruction aims to learn a 3D representation capable of synthesizing novel views from 2D images. This technology can reduce reliance on manual 3D modeling and has broad applicability across industries such as film production, robotic navigation, cultural heritage preservation, VR/AR and computer graphics. Recently, learning-based 3D reconstruction methods—such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2021)—have garnered considerable attention for their capacity to model continuous 3D scenes. Among these methods, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) stands out because of its point-based, explicit representation, which is compatible with con-

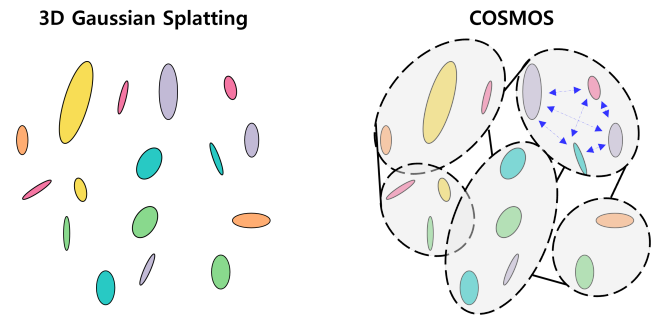


Figure 1: Comparison of the Optimization Process between 3DGS and COSMOS. While 3DGS learns each Gaussian independently to fit the training views, COSMOS leverages supergaussian grouping to inject spatial priors during optimization. Through inter-group learning, it captures the global 3D structural context, while intra-group regularization prevents Gaussians within similar geometric structures from diverging in different directions.

ventional 3D rendering pipelines and capable of high-quality novel view synthesis.

Despite its advantages in dense-view settings, optimizing 3DGS with sparse input views remains challenging. In practice, acquiring dense multi-view images for a single scene is often prohibitively costly and time-consuming, highlighting the need for models that remain robust under limited-view conditions. The primary reason for 3DGS’s limited generalization under sparse inputs is its reliance on simple pixel-wise photometric loss between rendered and training images without leveraging explicit 3D structural priors. As a result, 3DGS tends to overfit and fails to generalize to unseen viewpoints.

To address these limitations, recent studies have introduced prior knowledge to improve novel view synthesis under sparse input conditions (Mihajlovic et al. 2024; Xiong et al. 2023). A representative approach involves using pseudo-depth maps generated by auxiliary or pretrained models to regularize the geometry (Li et al. 2024; Deng et al. 2022; Xu et al. 2024; Chung, Oh, and Lee 2024; Zhu et al. 2024). While these approaches aim to preserve structural

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

consistency under sparse input conditions, their performance is highly dependent on the accuracy of the predicted depths, and the absolute scales of depth values vary across views, causing misalignment (Park, Ryu, and Kim 2025; Xu et al. 2025). To overcome these limitations, we propose an alternative 3D prior that provides structural cues without relying on depth supervision.

We propose **COherent SuperGaussian MOdeling with Spatial Priors (COSMOS)**, a framework that preserves the explicit, point-based representation of 3DGS while overcoming its limitations. Inspired by superpoints (Landrieu and Simonovsky 2018) in semantic segmentation, COSMOS redefines the grouping criteria to suit the 3DGS context. Specifically, it groups Gaussians into supergaussians based on local geometric relations and appearance features, thereby injecting structured 3D priors into the learning process.

These supergaussians serve two primary purposes. First, compressing thousands of Gaussians into only a few tens groups, COSMOS enables efficient global self-attention (Vaswani et al. 2017) across supergaussian groups, as well as distance-based sparse self-attention (Child et al. 2019) among individual Gaussians. This design allows the network to learn 3D feature vectors that capture both global and local spatial structure. These features guide Gaussian-attribute prediction and promote coherent 3D geometry under sparse input conditions. Second, intra-group regularization suppresses divergence among Gaussians belonging to the same structural region, thereby mitigating the occurrence of floaters. By injecting these supergaussian-based 3D priors into the model (Figure 1), COSMOS achieves robust and structurally coherent novel view synthesis. It also overcomes the sparse-view limitations of 3DGS.

In summary, our contributions are as follows:

1. We propose COSMOS, an attribute prediction framework that introduces a novel 3D structural prior through our newly defined inter-group supergaussian representation. COSMOS integrates global supergaussian embeddings with local Gaussian features to achieve structurally consistent predictions of Gaussian attributes.
2. Without relying on external depth supervision, COSMOS introduces an intra-group positional regularization to suppress floaters and enhance structural consistency under sparse-view conditions.
3. COSMOS achieves superior novel view synthesis performance under extremely sparse input conditions. Compared to recent state-of-the-art (SOTA) methods in 3D scene reconstruction, COSMOS demonstrates strong performance on the Blender and DTU datasets using as few as three input views. Experimental results show that our approach effectively reconstructs complex structures and high-frequency details even with minimal supervision.

2 Related Work

2.1 Neural Representations in 3D Reconstruction

3D reconstruction aims to synthesize images from unobserved camera viewpoints using a set of observed

views. Traditional approaches such as Structure-from-Motion (SfM) (Schonberger and Frahm 2016) and Multi-View Stereo (MVS) have long been studied but suffer from limitations including complex pipelines and fragile matching procedures. To overcome these challenges, recent research has increasingly focused on learning-based methods for novel view synthesis.

Among these, NeRF (Mildenhall et al. 2021) models a scene as an implicit continuous function that maps 3D coordinates and viewing directions to color and density, producing photorealistic novel views. However, its implicit formulation lacks explicit geometry for direct manipulation and requires expensive per-ray sampling and integration, making training slow (Li et al. 2023; Jambon et al. 2023).

In contrast, the recently introduced 3DGS (Kerbl et al. 2023) offers an explicit, point-based 3D representation that enables fast training and compatibility with traditional 3D rasterization pipelines. Each 3D Gaussian encodes position, covariance, color, and opacity, and is rendered as an ellipsoidal volume that can overlap with multiple pixels, leading to efficient and photorealistic rendering. Thanks to its simplicity and efficiency, 3DGS has rapidly become a central foundation for follow-up research, inspiring numerous extensions for memory efficient, dynamic scenes, and high-fidelity rendering (Hamdi et al. 2024; Lee et al. 2024; Sun et al. 2024; Wu et al. 2024; Yu et al. 2024).

Nevertheless, 3DGS still requires a large number of input views to achieve reliable novel view synthesis. When trained under sparse input conditions, it tends to fail to preserve object integrity and exhibits significant performance degradation in unseen views.

2.2 Few-Shot 3D Reconstruction

Few-shot 3D reconstruction, which aims to perform novel view synthesis from a limited set of input views, has been actively studied for years (Yu et al. 2021; Niemeyer et al. 2022; Xu et al. 2022, 2025; Somraj, Karanayil, and Soundararajan 2023; Yang et al. 2024; Johari, Lepoittevin, and Fleuret 2022). DietNeRF (Jain, Tancik, and Abbeel 2021) improves performance by injecting semantic information into the reconstruction process using CLIP (Radford et al. 2021), a large-scale image-text alignment model. FreeNeRF (Yang, Pavone, and Wang 2023) identifies the high-frequency components of NeRF’s positional encoding as a key contributor to overfitting. It improves generalization by partially masking these positional encodings during training. Other works have enhanced novel view synthesis by guiding the model with depth-based regularization to encourage more plausible geometry. This approach constrains the scene’s appearance using predicted depths (Li et al. 2024; Deng et al. 2022; Xu et al. 2024). However, such approaches rely on auxiliary or pretrained models to obtain depth supervision, limiting both flexibility and scalability.

In contrast, COSMOS instead regularizes the appearance without any depth supervision. This strategy mitigates overfitting and the occurrence of floaters, ultimately improving novel view synthesis under sparse input settings.

2.3 Structured Group Representations

The concept of superpoints originates from the idea of superpixels (Achanta et al. 2012) in 2D image processing, where geometrically homogeneous point sets are grouped into a single unit. This approach was first introduced in the context of 3D semantic segmentation. Superpointgraph (Landrieu and Simonovsky 2018) estimates the geometric properties of each superpoint by computing each superpoint’s coordinate covariance and analyzing its eigenvalues to derive descriptors (length, planarity, volumetricity, etc.) Based on these descriptors, the spatial relationships between superpoints are modeled as a graph, with edges encoding various geometric attributes such as relative position offsets, volume differences, and shape ratios. This representation allows for efficient processing of large-scale point clouds while enhancing structural expressiveness. Superpoints have since been widely adopted across various large-scale point cloud tasks, leading to a surge in related research (Robert, Ragué, and Landrieu 2023; Kolodiaznyy et al. 2024; Hui et al. 2021). COSMOS builds upon these prior works by incorporating the superpoint concept into few-shot 3DGS, introducing so-called supergaussians in 3DGS. Gaussians are grouped into supergaussians according to our defined grouping criterion based on local geometric and appearance features, injecting spatial bias and structural priors into the 3DGS framework. By leveraging these 3D priors, COSMOS achieves more robust generalization in few-shot 3D reconstruction scenarios.

3 Methodology

Our approach is built upon 3DGS, which enables explicit, point-based 3D reconstruction (Section 3.1). COSMOS begins by grouping 3D Gaussians into supergaussians based on superpoint-inspired grouping (Section 3.2). We then use a transformer architecture to compute global embeddings for each supergaussian, and local embeddings for individual Gaussians which are fed into the Gaussian attribute prediction module (Section 3.3). Finally, the model is optimized using a supergaussian-level regularization loss (Section 3.4). Figure 2 illustrates the overall architecture of COSMOS.

3.1 Preliminaries for 3DGS

3D Gaussians are point-based primitives used to represent a 3D scene, where each Gaussian is parameterized by its position, covariance, color, and opacity. A Gaussian can be defined by a center position x and a 3D covariance matrix Σ in 3D space:

$$G(x) = e^{-\frac{1}{2}(x)^T \Sigma^{-1} x} \quad (1)$$

Since the covariance matrix must be positive definite, direct optimization via gradient descent can easily result in invalid matrices. To ensure valid and stable optimization, each Gaussian’s covariance is reparameterized as a combination of a rotation matrix and a scaling matrix. Specifically, a 3D scale vector S encodes the axis-aligned scaling and a unit-norm quaternion encodes the rotation R .

$$\Sigma = R S S^T R^T \quad (2)$$

After optimization, we project the Gaussians into camera space using the viewing transformation W . Then we apply J , the Jacobian of the projective transformation’s affine approximation, to the Gaussian parameters to enable rendering.

$$\Sigma' = J W \Sigma W^T J^T \quad (3)$$

During rendering, each Gaussian is rasterized and contributes to multiple pixels along the ray. The final pixel color C is computed by aggregating the projected attributes of N Gaussians along each ray:

$$C = \sum_{i \in N} \left(\alpha_i c_i \prod_{j=1}^{i-1} (1 - \alpha_j) \right) \quad (4)$$

where c_i denotes the color of each point, and α_i is computed by evaluating the 2D Gaussian obtained from projecting the 3D Gaussian with covariance Σ onto the image plane, and scaling it by a learned per-Gaussian opacity parameter. 3DGS supervises the model with an \mathcal{L}_1 loss (and an $\mathcal{L}_{\text{ssim}}$) between the rendered image and the ground-truth image. The Gaussian attributes are optimized to minimize this discrepancy, improving scene fidelity.

3.2 Supergaussian Grouping

To alleviate floaters and structural collapse in sparse-view settings, we begin by grouping Gaussians to introduce 3D priors into 3DGS. Specifically, we compute local geometric descriptors for each Gaussian, including linearity (ℓ_i), scattering (s_i), verticality (v_i), and planarity (p_i) (Landrieu and Simonovsky 2018). These descriptors are concatenated with the Gaussian’s position (x_i), color (c_i), and scale (σ_i) to form the feature vector used for grouping. We then apply the ℓ_0 -Cut pursuit algorithm (Landrieu and Obozinski 2017) to partition the Gaussians into groups. This procedure requires only a few graph-cut iterations and efficiently produces supergaussians.

$$f_i = [x_i, c_i, \sigma_i, \ell_i, s_i, v_i, p_i] \quad (5)$$

Here, i denotes the index of a Gaussian. Unlike original superpoints, COSMOS additionally incorporates Gaussian scale as a grouping feature. The scale of a Gaussian plays a critical role in the reconstruction process: smaller Gaussians tend to capture fine-grained details and high-frequency regions, whereas larger Gaussians contribute to the reconstruction of coarser global structures. By including scale in the feature representation, the model can distinguish whether a given supergaussian group is intended to recover local detail or overall shape.

$$S_g = \{i \mid k_i = g\}, \quad g = 0, \dots, G - 1 \quad (6)$$

S_g denotes the set of Gaussian indices i belonging to supergaussian g , where G is the total number of supergaussian groups and k_i indicates the group assignment of Gaussian i . Following (Landrieu and Simonovsky 2018), we also compute each group’s spatial covariance and derive group-level linearity, planarity, verticality, and scattering descriptors to characterize its geometry.

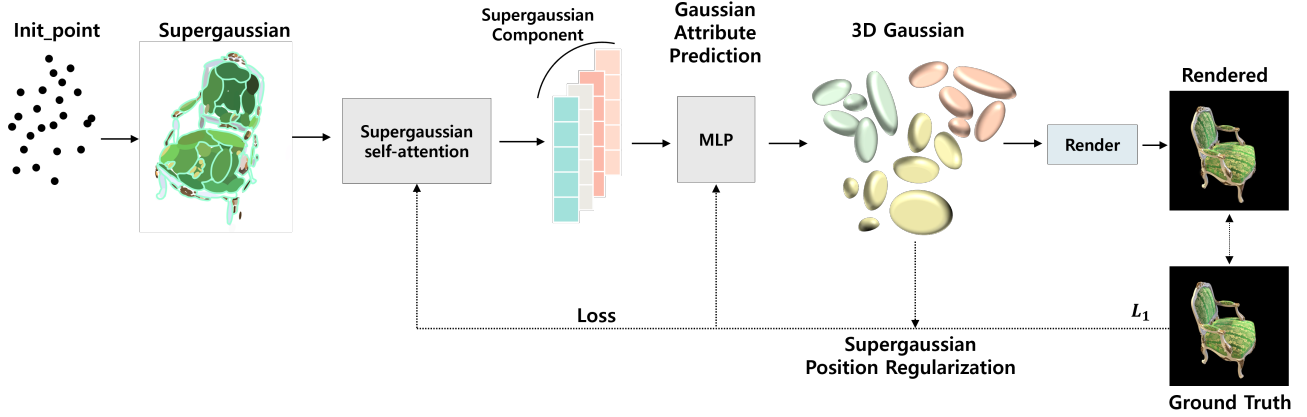


Figure 2: Architecture of COSMOS. To incorporate 3D priors, we initially group Gaussians into supergaussians based on geometric cues. A self-attention module operating at the supergaussian level is then employed to generate 3D feature vectors, enabling each Gaussian to be trained in a structurally informed manner rather than independently. Furthermore, we introduce a regularization term that constrains the positional deviation of Gaussians within the same supergaussian, which helps suppress floaters and alleviates structural collapse under sparse input conditions.

3.3 Supergaussian Self-Attention

To inject spatial bias into the representation, self-attention can be effectively leveraged for point cloud embedding (Guo et al. 2021; Yu et al. 2022). However, since a typical 3DGS scene is composed of tens to hundreds of thousands of Gaussians, directly applying global attention across all Gaussians is computationally prohibitive. In contrast, COSMOS addresses this limitation by grouping large numbers of Gaussians into a few tens supergaussians, thus enabling efficient global self-attention at the group level.

We first perform group-wise max pooling to extract a representative feature s_g for each supergaussian, using the previously computed group assignments and per-Gaussian features.

$$s_g = \max_{i \in S_g} f_i \quad \text{and} \quad s_g = [x_g; r_g] \quad (7)$$

Positional encoding γ (Mildenhall et al. 2021) is then applied to the group center position x_g which we separate from the remaining features r_g . After applying positional encoding to x_g , it is concatenated with r_g , and the combined features are passed through a linear projection layer W_{in} , b_{in} to adjust their dimensionality.

$$u_g = W_{\text{in}} \text{Norm}([\gamma(x_g); r_g]) + b_{\text{in}} \quad (8)$$

The features are then normalized and fed through linear projection layers to obtain the query, key, and value embeddings for self-attention.

$$q_g^h = W_h^Q u_g, \quad k_g^h = W_h^K u_g, \quad v_g^h = W_h^V u_g \quad (9)$$

Here, W_h^Q , W_h^K , and $W_h^V \in \mathbf{R}^{d_{\text{model}} \times d_h}$ are learned projection matrices that map the input features to queries q_g^h ,

keys k_g^h , and values v_g^h , respectively. Using these queries, keys, and values, we compute scaled dot-product attention across all supergaussians. The resulting attention weights capture the relevance between supergaussians, and the final embedding encodes the inter-group spatial context for each group. The learned inter-group spatial features are then passed through an activation function and a final output projection layer, yielding a set of supergaussian embeddings that capture global structural context.

Aiming to capture high-frequency details and fine-grained spatial variations that cannot be fully learned through group-based dense attention, we introduce a sparse local attention mechanism (Child et al. 2019) in addition to global group-level attention. In this design, each Gaussian is represented by a feature vector encoding both appearance and geometric attributes, and its 3D spatial coordinates are positionally encoded. Each Gaussian then attends to its ten nearest neighbors using a sparse transformer. Finally, the outputs from the global group attention and local sparse attention are concatenated to form a unified 3D feature vector for each Gaussian that integrates both global and local information.

To predict the final Gaussian attributes—namely position, orientation, scale, color, and opacity—COSMOS employs separate Residual MLPs (Mihajlovic et al. 2023), each dedicated to one specific attribute. These MLPs take as input the concatenation of the Gaussian’s spatial coordinates and the corresponding unified 3D feature, enabling structurally aware attribute prediction.

3.4 Supergaussian Position Regularization

To mitigate the tendency of Gaussians to overfit to training views, COSMOS introduces a supergaussian regularization that enforces intra-group attribute consistency. This regularization encourages Gaussians within the same group to learn coherent representations, improving generalization to unseen views.

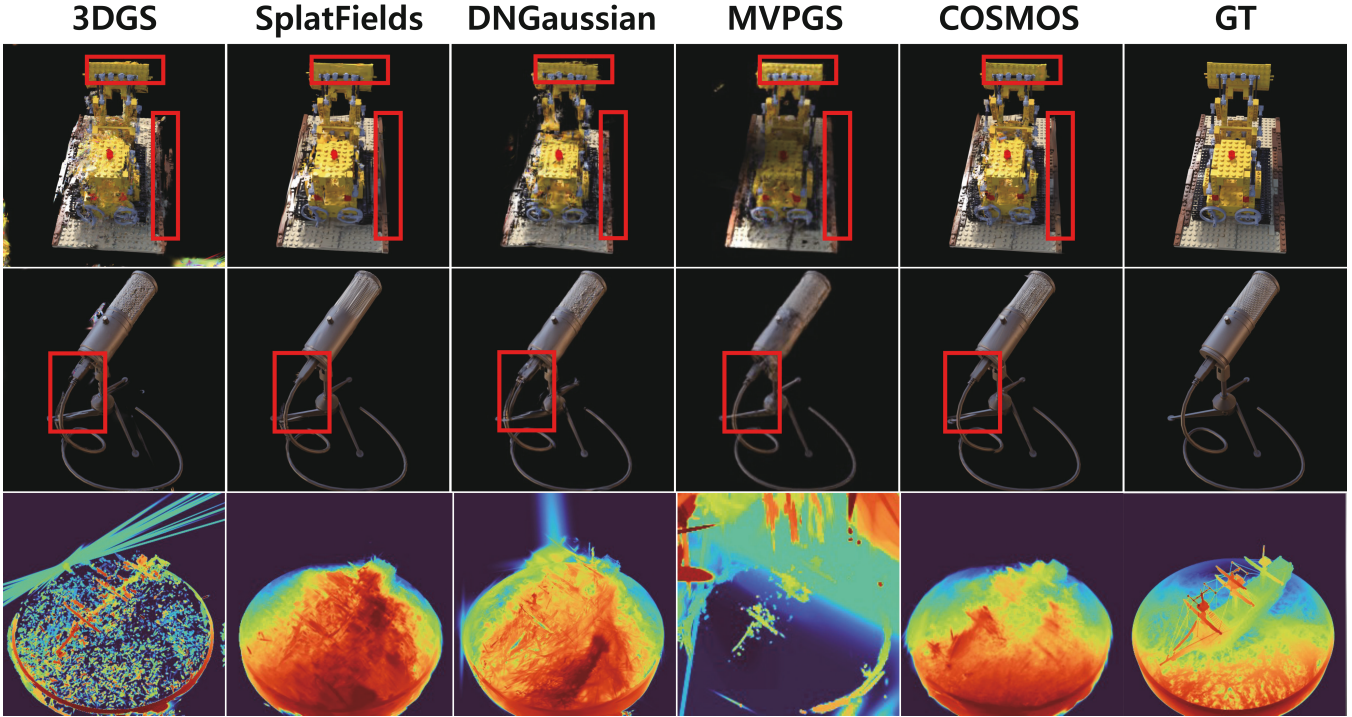


Figure 3: Qualitative Comparison on Blender. We train COSMOS and recent competitive models with 3 input views and render novel views for evaluation. COSMOS effectively suppresses floaters and restores fine high-frequency details (see red boxes). The third-row depth maps show that COSMOS reconstructs the smoothest and most continuous geometry without any depth supervision.

Supergaussian position regularization consists of two regularization terms. First, it constrains the learning of Gaussian positions to prevent them from drifting too far from their neighboring Gaussians.

$$D_{\text{avg}} = \frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{N}(i)| + \varepsilon} \sum_{j \in \mathcal{N}(i)} \|x_i - x_j\|_2 \quad (10)$$

$\mathcal{N}(i) = \{j \mid (i, j) \in \mathcal{E}\}$ denotes the set of neighboring Gaussians of the i -th Gaussian, where \mathcal{E} is the set of edges connecting neighboring Gaussians, and N is the total number of Gaussians.

The second term encourages each Gaussian to remain close to the center of its assigned supergaussian, preventing Gaussians within similar geometric structures from drifting in different directions. This promotes continuity and structural consistency during learning.

$$D_{\text{ctr}} = \frac{1}{N} \sum_{i=1}^N \left\| x_i - \frac{1}{|S_{k_i}| + \varepsilon} \sum_{j \in S_{k_i}} x_j \right\|_2^2 \quad (11)$$

$|S_{k_i}|$ represents the size of the supergaussian to which the i -th Gaussian belongs. The overall position regularization loss is formulated as follows:

$$\mathcal{L}_{\text{pos}} = \lambda_1 D_{\text{avg}} + \lambda_2 D_{\text{ctr}} \quad (12)$$

These constraints mitigate overfitting under sparse-view training, preventing Gaussians from drifting to spurious positions (which cause ‘floaters’ artifacts).

The final loss combines the image reconstruction term \mathcal{L}_1 (with $\mathcal{L}_{\text{ssim}}$), the regularization term \mathcal{L}_{pos} and an optional mask-based loss $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_1 + (1 - \lambda) \mathcal{L}_{\text{SSIM}} + \lambda_{\text{pos}} \mathcal{L}_{\text{pos}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} \quad (13)$$

Here, λ balances \mathcal{L}_1 and $\mathcal{L}_{\text{ssim}}$. λ_{pos} and λ_{mask} weight the regularization terms.

4 Experiments

4.1 Setup

Datasets and Metrics We conduct experiments on the Blender dataset (Mildenhall et al. 2021) provided by NeRF, and the real-world multi-view dataset DTU (Jensen et al. 2014). For Blender, we use two settings: using 3 or 6 training views, following the same view sampling protocol as previous works (Mihajlovic et al. 2024). Evaluation is then performed on 200 test views, as in the original NeRF setting. For DTU, we adopt the same experimental setup as prior studies (Huang et al. 2024; Niemeyer et al. 2022), using 3 training views for each scene. We report PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) metrics, following standard evaluation protocols used in prior novel view synthesis studies.

Method	$N = 3$			$N = 6$		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DietNeRF	10.47	0.57	0.42	14.10	0.66	0.34
FreeNeRF	16.75	0.72	0.23	23.67	0.87	0.11
MVPGS	16.64	0.66	0.30	21.07	0.78	0.17
DNGaussian	16.28	0.74	0.21	21.79	0.85	0.11
SplatFields	<u>18.46</u>	<u>0.77</u>	<u>0.20</u>	22.26	<u>0.86</u>	<u>0.15</u>
3DGS	14.77	0.58	0.32	21.05	0.80	0.16
COSMOS (Ours)	18.96	0.79	0.19	<u>22.78</u>	0.85	<u>0.15</u>

Table 1: Quantitative Comparison on Blender. We mark the **best** and second best results.

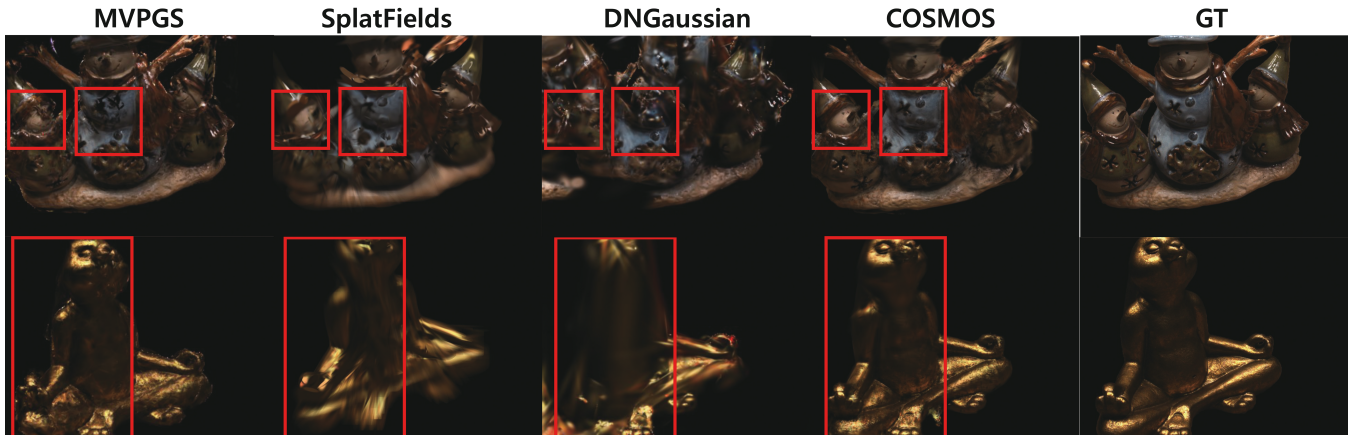


Figure 4: Qualitative Comparison on DTU. We train MVPGS, SplatFields, DNGaussian, and COSMOS using 3 input views and render novel views for evaluation. COSMOS demonstrates superior novel view synthesis performance, even on real-world datasets with complex textures and structures.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
FreeNeRF	10.10	0.24	0.63
MVPGS	20.71	0.75	<u>0.19</u>
DNGaussian	16.07	0.80	0.15
SplatFields	<u>21.07</u>	<u>0.78</u>	0.20
3DGS	19.40	0.54	0.48
COSMOS (Ours)	21.86	0.76	0.23

Table 2: Quantitative Comparison on DTU ($N = 3$). We mark the **best** and second best results.

Implementation details All experiments are performed on a single NVIDIA A6000 GPU. We build our implementation on SplatFields (Mihajlovic et al. 2024), removing its encoder component. We train each model for 12,000 iterations. Grouping into supergaussians is performed after the first 100 iterations of training, once the model has learned a coarse representation of the scene. During the subsequent densification stage, where Gaussians may be split or cloned, newly created Gaussians inherit the group identity of their origin Gaussian to preserve grouping consistency. We treat the following as hyperparameters: (i) the number of neighbors for computing each Gaussian’s geometric descriptors (l_i, s_i, v_i, p_i , etc.), and (ii) the grouping strength threshold in

Cut Pursuit. These can be adjusted per dataset or reconstruction setting. We found that the effective number of Super-Gaussian groups is about 0.02%–0.08% of the initial number of Gaussians in a scene (for these datasets), corresponding to approximately 16–64 groups. We use $\lambda_1 = \lambda_2 = 1$ for the two regularization terms, and set the SSIM blending factor $\lambda = 0.8$ (following prior work). We set $\lambda_{pos} = 0.2$. A mask loss is used for Blender/DTU (background masking), with $\lambda_{mask} = 0.1$.

Comparing methods We compare our method against 3DGS (Kerbl et al. 2023), its sparse-view variant (Xu et al. 2024; Li et al. 2024; Mihajlovic et al. 2024), and SOTA NeRF-based methods designed for sparse input settings (Yang, Pavone, and Wang 2023; Jain, Tancik, and Abbeel 2021). For evaluation, we adopt the metrics reported in the original papers when available; otherwise, we use the official implementations to compute the performance.

4.2 Comparison

Blender Qualitative and quantitative results on the Blender dataset using 3 and 6 input views are presented in Table 1 and Figure 3. COSMOS achieves the best performance across all metrics in the highly sparse 3-view setting, outperforming recent SOTA methods. In particular, the depth map renderings demonstrate that COSMOS constructs

Method	MAE ↓	SROCC ↑
DNGaussian	0.53	0.12
MVPGS	0.28	0.44
COSMOS (Ours)	0.22	0.41

Table 3: Evaluation of depth maps from two depth-based comparison models on the DTU dataset with three input views. The reference depth is obtained using 3DGS trained with dense input views.

continuous geometry and effectively suppresses floaters and noise without relying on any external depth priors. Moreover, even with only 6 input views, COSMOS achieves competitive performance with recent SOTA methods, confirming that the supergaussian-based 3D priors are effective even under sparse inputs.

DTU Qualitative results and visualizations using 3 input views on the dataset are presented in Table 2 and Figure 4. COSMOS achieves the best performance in terms of PSNR and shows competitive results in SSIM compared to recent methods. Although DNGaussian attains the best SSIM and LPIPS values, our visual results show that COSMOS reconstructs more complex, high-frequency structures without introducing floaters. These results indicate that learning with 3D priors via Gaussian grouping effectively preserves fine-grained geometric details.

Following prior work (Somraj and Soundararajan 2023; Somraj, Karanayil, and Soundararajan 2023), we evaluate absolute depth prediction performance using the mean absolute error (MAE) of per-pixel depth. In addition, we assess relative depth prediction performance using the Spearman Rank-Order Correlation Coefficient (SROCC) (Corder and Foreman 2014), which measures the correlation between predicted and ground-truth pixel depths. The predicted depths of 3DGS trained with dense input views are used as pseudo-ground truth. As shown in Table 3, COSMOS achieves the lowest MAE and competitive SROCC. These results indicate that, despite not relying on depth supervision, COSMOS effectively predicts depth compared to depth-based methods such as DNGaussian and MVPGS, demonstrating the benefit of incorporating supergaussian-based 3D priors.

4.3 Ablation Study

In this section, we perform an ablation study on the 3-view Blender dataset to analyze the effectiveness of COSMOS. As shown in Table 4 and Figure 5, removing the supergaussian self-attention module leads to significant performance degradation, demonstrating that learning global inter-group features is crucial for producing structurally coherent Gaussian predictions. Similarly, removing the supergaussian position regularization results in the emergence of noise and floaters around object boundaries, which our method effectively suppresses. Overall, these ablations confirm that both the supergaussian grouping/self-attention and the positional regularization are critical for robust structure reconstruction under sparse views.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o SG Self-Attention	18.74	0.77	0.20
w/o SG Regularization	18.24	0.76	0.21
w/o SG Regul + SG SA	18.22	0.76	0.21
Full model	18.96	0.79	0.19

Table 4: Quantitative Ablation Study of COSMOS. SG denotes supergaussian.

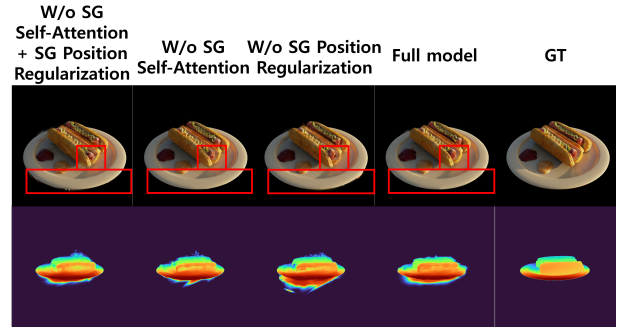


Figure 5: Qualitative Ablation Study of COSMOS. The red boxes highlight that the Full model captures the most detailed textures. In the depth map renderings, the object surfaces show that the Full model maintains the most consistent and continuous geometry.

4.4 Limitation

Setting the optimal number of supergaussians may require cross-validation. An improper choice could under-segment or over-segment the scene. In future work, an adaptive grouping strategy could alleviate this sensitivity to the grouping parameter.

5 Conclusion

In this work, we proposed COSMOS, a novel framework that improves 3DGS under sparse-view conditions by grouping Gaussians into supergaussians based on local geometric and appearance cues. We design a representation learning mechanism that integrates both global and local spatial information, enabling structure-aware prediction of Gaussian attributes (position, rotation, etc.). Furthermore, intra-group positional regularization is applied to suppress floaters and maintain structural coherence. Extensive experiments on the Blender and DTU benchmarks demonstrate that COSMOS consistently outperforms existing methods in few-shot 3D reconstruction.

The idea of supergaussian-based grouping opens up avenues beyond novel view synthesis — for instance, interactive 3D editing or semantic segmentation of Gaussian splats could benefit from our structured priors. We hope this work encourages further exploration of structured priors in point-based 3D reconstruction.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00459618, Research on improving intelligent command and control capabilities based on generative AI and real-time 3D digital twin construction) and Next-Generation Intelligent Patrol Platform (Mobile police station) Program through the Korea Institutes of Police Technology (KIPoT) funded by the Korean National Police Agency. (No. RS-2025-25393280)

References

- Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.
- Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Chung, J.; Oh, J.; and Lee, K. M. 2024. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.
- Corder, G. W.; and Foreman, D. I. 2014. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.
- Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12882–12891.
- Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; and Hu, S.-M. 2021. Pct: Point cloud transformer. *Computational visual media*, 7(2): 187–199.
- Hamdi, A.; Melas-Kyriazi, L.; Mai, J.; Qian, G.; Liu, R.; Vondrick, C.; Ghanem, B.; and Vedaldi, A. 2024. Ges: Generalized exponential splatting for efficient radiance field rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19812–19822.
- Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, 1–11.
- Hui, L.; Yuan, J.; Cheng, M.; Xie, J.; Zhang, X.; and Yang, J. 2021. Superpoint network for point cloud oversegmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5510–5519.
- Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5885–5894.
- Jambon, C.; Kerbl, B.; Kopanas, G.; Diolatzis, S.; Leimkühler, T.; and Drettakis, G. 2023. Nerfshop: Interactive editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1).
- Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 406–413.
- Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18365–18375.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.
- Kolodiazhnyi, M.; Vorontsova, A.; Konushin, A.; and Rukhovich, D. 2024. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20943–20953.
- Landrieu, L.; and Obozinski, G. 2017. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4): 1724–1766.
- Landrieu, L.; and Simonovsky, M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4558–4567.
- Lee, J. C.; Rho, D.; Sun, X.; Ko, J. H.; and Park, E. 2024. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21719–21728.
- Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. 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*, 20775–20785.
- Li, R.; Gao, H.; Tancik, M.; and Kanazawa, A. 2023. Nerfacc: Efficient sampling accelerates nerfs. In *Proceedings of the IEEE/CVF international conference on computer vision*, 18537–18546.
- Mihajlovic, M.; Prokudin, S.; Pollefeys, M.; and Tang, S. 2023. Resfields: Residual neural fields for spatiotemporal signals. *arXiv preprint arXiv:2309.03160*.
- Mihajlovic, M.; Prokudin, S.; Tang, S.; Maier, R.; Bogo, F.; Tung, T.; and Boyer, E. 2024. Splatfields: Neural gaussian splats for sparse 3d and 4d reconstruction. In *European Conference on Computer Vision*, 313–332. Springer.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5480–5490.
- Park, H.; Ryu, G.; and Kim, W. 2025. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21600–21609.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Robert, D.; Raguét, H.; and Landrieu, L. 2023. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17195–17204.
- Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.
- Somraj, N.; Karanayil, A.; and Soundararajan, R. 2023. Simplenerf: Regularizing sparse input neural radiance fields with simpler solutions. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.
- Somraj, N.; and Soundararajan, R. 2023. Vip-nerf: Visibility prior for sparse input neural radiance fields. In *ACM SIGGRAPH 2023 conference proceedings*, 1–11.
- Sun, J.; Jiao, H.; Li, G.; Zhang, Z.; Zhao, L.; and Xing, W. 2024. 3dstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20675–20685.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.
- Xiong, H.; Muttukuru, S.; Upadhyay, R.; Chari, P.; and Kadambi, A. 2023. Sparsegs: Real-time 360 $\{\backslash\deg\}$ sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*.
- Xu, D.; Jiang, Y.; Wang, P.; Fan, Z.; Shi, H.; and Wang, Z. 2022. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, 736–753. Springer.
- Xu, W.; Gao, H.; Shen, S.; Peng, R.; Jiao, J.; and Wang, R. 2024. Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. In *European Conference on Computer Vision*, 203–220. Springer.
- Xu, Y.; Wang, L.; Chen, M.; Ao, S.; Li, L.; and Guo, Y. 2025. DropoutGS: Dropping Out Gaussians for Better Sparse-view Rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 701–710.
- Yang, C.; Li, S.; Fang, J.; Liang, R.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2024. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *arXiv preprint arXiv:2402.10259*.
- Yang, J.; Pavone, M.; and Wang, Y. 2023. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8254–8263.
- Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4578–4587.
- Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19313–19322.
- Yu, Z.; Chen, A.; Huang, B.; Sattler, T.; and Geiger, A. 2024. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19447–19456.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhu, Z.; Fan, Z.; Jiang, Y.; and Wang, Z. 2024. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, 145–163. Springer.