

VISTA-GS: MVS-Guided Virtual View Augmentation for Sparse-View 3D Gaussian Splatting

Hongsheng Huang¹, Yaxin Li^{1,4,*}, Shengjun Tang², Siqi Du³, Mostafa Mahmoud¹, Mahmoud Adham¹, Wu Chen¹

¹ Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, P.R. China

² Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, P.R. China

³ College of Urban and Environmental Sciences, Peking University, Beijing, P.R. China

⁴ Micro Dimension Technology Limited, Hong Kong, P.R. China

Keywords: 3D Gaussian Splatting, Sparse-View Novel View Synthesis, Appearance Overfitting, Virtual View Augmentation, Out-of-Distribution Generalization.

Abstract

3D Gaussian Splatting (3DGS) has emerged as a leading technique for novel view synthesis (NVS), yet its performance degrades drastically under sparse-view conditions. While existing methods have sought to address this by incorporating accurate 3D geometry via Multi-View Stereo (MVS) or LiDAR priors, the view-dependent appearance parameters (i.e., spherical harmonics) remain exclusively optimized on the limited training views, leading to severe appearance overfitting. This is the fundamental reason why these geometry-enhanced methods still fail to generalize to out-of-distribution (OOD) viewpoints with large baselines, such as lane-changing trajectories in autonomous driving. To address this limitation, we propose VISTA-GS (Virtual Image Synthesis and Training Augmentation), a framework that synergizes MVS-based dense initialization with a physically-grounded virtual view augmentation strategy. Specifically, we position virtual cameras at strategic offsets around the original viewpoints and render virtual training images with binary validity masks via alpha-blending. By computing photometric losses exclusively within valid mask regions, VISTA-GS injects explicit angular constraints into the optimization process, effectively regularizing view-dependent appearance without relying on any external generative model. Experiments on the LLFF benchmark and a real-world LiDAR-scanned dataset demonstrate that our method achieves state-of-the-art NVS quality under sparse-view settings, with particularly significant improvements on challenging OOD viewpoints.

1. Introduction

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has emerged as a leading technique for novel view synthesis (NVS), offering real-time, high-fidelity rendering through explicit Gaussian primitives. However, its performance critically depends on dense multi-view constraints and reliable Structure-from-Motion (SfM) (Schonberger and Frahm, 2016) initialization. In single-trajectory scanning scenarios such as autonomous driving, onboard cameras typically capture only a sparse set of views along a fixed ego-vehicle trajectory. Under such severely ill-posed conditions, 3DGS tends to overfit the narrow training viewpoints. As a result, when rendering out-of-distribution (OOD) perspectives with large baselines—such as lane-changing trajectories—the synthesis quality collapses, producing pervasive floaters and blurry artifacts that are unacceptable for safety-critical applications.

Existing methods for sparse-view 3DGS can be broadly grouped by the type of prior information they exploit, yet none adequately addresses the lane-changing rendering problem. One line of work incorporates external 2D priors to regularize scene geometry. Methods such as DN-Gaussian (Li et al., 2024) and D²GS (Song et al., 2025) leverage monocular depth estimators, whereas LM-Gaussian (Yu et al., 2024) and See in Detail (He et al., 2025) utilize semantic features from large foundation models. Although these 2D priors provide valuable structural hints, the depth maps produced by monocular models inevitably suffer from inherent scale ambiguities and multi-view inconsistencies, which can misguide the 3DGS optimization. Closely related are methods that employ generative models to hallucinate unobserved viewpoints, such as ReconDreamer (Ni et al., 2025) and DIFix3D+ (Wu et al., 2025). While these ap-

proaches can produce visually plausible results, they rely on large-scale pretrained models and require high-quality initial inputs. Moreover, the outputs of such models often lack strict geometric fidelity—a critical limitation in precision-demanding driving scenarios.

Another category of methods introduces explicit 3D priors, such as Multi-View Stereo (MVS) or LiDAR point clouds, to provide geometry initialization and depth supervision. Representative methods include MVSGS (Fei et al., 2025), MVPGS (Xu et al., 2024) and LOD-3DGS (Cui et al., 2024). By grounding the Gaussian positions in reconstructed 3D structure, these approaches substantially improve geometric accuracy. However, a deeper issue remains: accurate geometry alone does not guarantee high-quality rendering from novel viewpoints. Since the view-dependent appearance parameters (i.e., spherical harmonics) of each Gaussian are still optimized exclusively on the limited training images, they inevitably overfit to the observed viewing angles. This appearance overfitting is the fundamental reason why even geometry-enhanced methods degrade under large viewpoint deviations. In essence, the core unsolved problem across existing approaches is the absence of explicit angular constraints on view-dependent appearance optimization.

To address this fundamental limitation, we propose VISTA-GS (Virtual Image Synthesis and Training Augmentation), a framework that synergizes MVS-based geometric reconstruction with a physically-grounded virtual view augmentation strategy. Rather than depending on external generative models or using 3D priors solely for initialization, our key insight is to exploit the geometric consistency already embedded in MVS point clouds to directly construct the missing angular

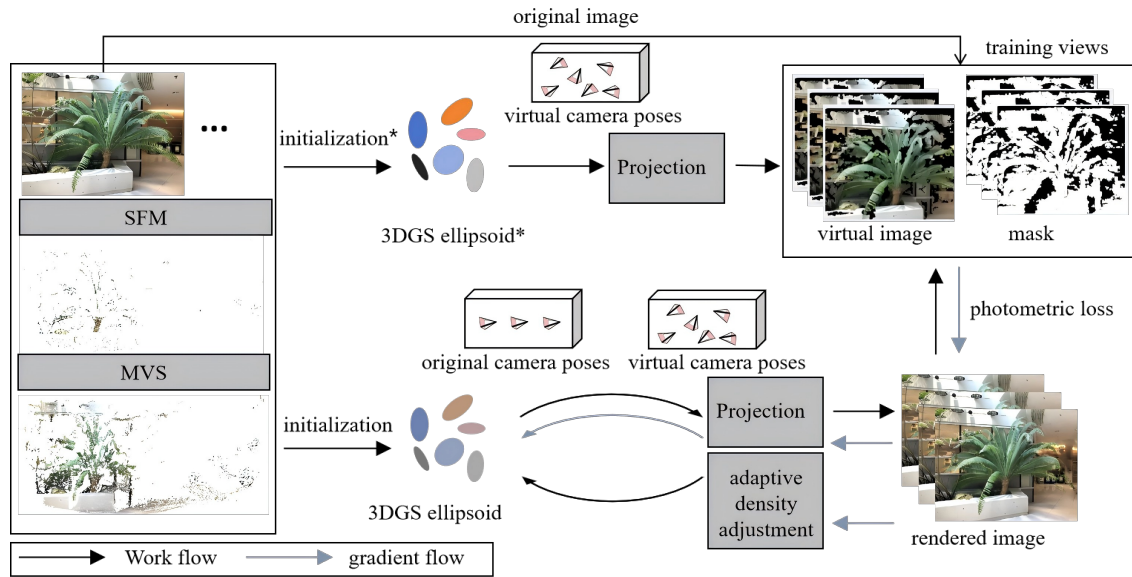


Figure 1. The framework of VISTA-GS.

supervision. Specifically, prior to the main 3DGS optimization, we position virtual cameras at strategic offsets around the original viewpoints and render virtual training images together with binary validity masks via alpha-blending. These masked virtual views augment the training set with diverse angular observations. By computing photometric losses exclusively within valid mask regions, our method effectively regularizes the spherical harmonics optimization across a broad range of viewing directions, preventing appearance overfitting while strictly avoiding the introduction of generative artifacts. Our main contributions are summarized as follows:

- We identify appearance overfitting—the exclusive optimization of view-dependent parameters on narrow training angles—as the root cause of rendering degradation under large viewpoint shifts, and propose VISTA-GS, a principled framework that resolves this issue through MVS-augmented angular supervision.
- We introduce a rendering-based virtual view generation mechanism with masked photometric supervision. This physically-grounded strategy provides explicit angular constraints without relying on any external generative model, offering a lightweight yet highly effective solution to the underconstrained sparse-view optimization.
- Extensive experiments on standard sparse-view benchmarks and a real-world LiDAR-scanned dataset demonstrate that VISTA-GS achieves state-of-the-art NVS quality, with particularly significant gains on challenging OOD viewpoints such as lane-changing trajectories.

2. Methodology

2.1 Overview

As illustrated in Figure 1, our pipeline is designed to process extremely sparse-view input data. Initially, we utilize COLMAP’s SfM and MVS modules to extract camera parameters and reconstruct a dense colored point cloud. To overcome the limitations of sparse SfM initialization, these MVS-derived

point clouds are initialized as opaque Gaussian ellipsoids. Subsequently, we project these ellipsoids using alpha-blending to render color images and corresponding binary masks at strategically defined virtual camera poses surrounding the original training views. These masked virtual images are then integrated into the training dataset. During the 3DGS optimization phase, photometric losses are computed exclusively within the valid mask regions. This augmentation strategy provides robust angular constraints, yielding reconstructions with superior geometric accuracy and enhanced generalization capabilities for novel view synthesis.

2.2 Preliminaries

The 3D Gaussian Splatting framework represents scenes through collections of anisotropic 3D Gaussians, each parameterized by position $\mu \in \mathbb{R}^3$, covariance matrix Σ , opacity α , and color c . The differentiable rendering employs alpha-blending along viewing rays:

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

where $C(\mathbf{r})$ denotes the rendered color for ray $\mathbf{r}$, and N represents the set of Gaussians intersecting the ray, ordered by depth.

Building upon this rendering formulation, our method augments the training set with virtual views synthesized from a frozen proxy Gaussian model constructed from MVS priors (detailed in Sec. 2.3). Since these virtual images may contain regions that lack valid geometric support, directly applying photometric supervision over the entire image would introduce erroneous gradients from unsupported areas. To address this, we introduce a masked photometric supervision strategy. Specifically, let $M(\mathbf{p}) \in \{0, 1\}$ denote a binary validity mask indicating whether pixel $\mathbf{p}$ is covered by sufficient MVS geometry; its construction is presented in Sec. 2.3. The photometric loss for virtual views is then computed exclusively over valid

regions:

$$\mathcal{L}_{\text{color}} = \frac{1}{|\mathcal{M}|} \sum_{\mathbf{p} \in \mathcal{M}} \left(\lambda_1 |C(\mathbf{p}) - \hat{C}(\mathbf{p})| + \lambda_2 \text{D-SSIM}(C(\mathbf{p}), \hat{C}(\mathbf{p})) \right) \quad (2)$$

where $\mathcal{M} = \{\mathbf{p} : M(\mathbf{p}) = 1\}$ denotes the set of valid pixels, $C(\mathbf{p})$ is the color rendered by the optimized 3DGS model, and $\hat{C}(\mathbf{p})$ is the corresponding target color from the proxy-rendered virtual view. This masked formulation ensures that the optimization receives supervision only from geometrically reliable regions, preventing the model from being misled by artifacts in unsupported areas.

2.3 MVS-Guided Virtual Image Generation

In the sparse-view setting, the limited number of input images provides insufficient multi-view constraints, causing the 3D Gaussians to easily overfit to the training views and generate severe floaters in unobserved regions. To inject robust geometric regularization into the optimization process, we propose to synthesize high-quality virtual images and corresponding validity masks from Multi-View Stereo (MVS) priors. These virtual observations serve as additional training samples that provide diverse angular supervision, guiding both the geometry and appearance of the Gaussians toward a globally consistent solution.

Extracting Clean MVS Priors. To compensate for the lack of geometric constraints, we utilize an off-the-shelf MVS pipeline to reconstruct a dense colored point cloud from the sparse training images. Given that raw MVS outputs inevitably contain noise and outliers, a Statistical Outlier Removal (SOR) filter is applied to obtain a clean and reliable point cloud. This filtered point cloud is denoted as $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^3$ represents the 3D spatial coordinate and $\mathbf{c}_i \in \mathbb{R}^3$ denotes the associated RGB color.

Constructing the Proxy Gaussian Model. Rather than directly optimizing this point cloud, we initialize a proxy 3D Gaussian Splatting model to render the MVS priors into 2D image space. For each point i , the Gaussian mean μ_i is anchored at $\mathbf{x}_i$, and the spherical harmonics (SH) coefficients are initialized directly from the color $\mathbf{c}_i$. To ensure that the proxy model faithfully reflects the raw MVS geometry without introducing optimization artifacts, all parameters of these Gaussians are frozen during rendering. We enforce an isotropic shape with a fixed small scale $s = 0.01$ and set the base opacity to be fully opaque ($\alpha_i = 1.0$). Consequently, the 3D covariance matrix Σ_i for each isotropic Gaussian is formulated as:

$$\Sigma_i = s^2 \mathbf{I} \quad (3)$$

where $\mathbf{I}$ is the 3×3 identity matrix.

Synthesizing Virtual Views and Masks. With the proxy model established, we project these frozen Gaussians onto pre-defined virtual camera poses. For a given pixel $\mathbf{p}$ on the virtual image plane, the overlapping Gaussians are sorted by depth, and the rendered color $C(\mathbf{p})$ is computed using standard α -blending:

$$C(\mathbf{p}) = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \quad (4)$$

where α'_i denotes the 2D projected opacity of the i -th Gaussian evaluated at pixel $\mathbf{p}$. Simultaneously, to distinguish regions with valid MVS geometric coverage from empty space, we

compute a per-pixel accumulated opacity following the same alpha-compositing principle:

$$M(\mathbf{p}) = \sum_{i \in \mathcal{N}} \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \quad (5)$$

Here, $M(\mathbf{p})$ represents the total opacity accumulated along the viewing ray at pixel $\mathbf{p}$. Since all proxy Gaussians are initialized with full base opacity ($\alpha_i = 1.0$), $M(\mathbf{p})$ naturally approaches 1 wherever MVS points project onto the image plane, and remains close to 0 in regions devoid of geometric observations. This near-binary distribution provides an effective indicator of geometric validity: pixels with $M(\mathbf{p}) \approx 1$ correspond to well-covered areas suitable for supervision, whereas pixels with $M(\mathbf{p}) \approx 0$ correspond to empty regions that should be excluded. We binarize $M(\mathbf{p})$ with a small threshold τ to obtain the final crisp validity mask, which defines the valid pixel set $\mathcal{M}$ used in the masked photometric loss (Eq. 2).

The resulting virtual images and validity masks are integrated into the training set alongside the original views. During optimization, this masked supervision ensures that the 3DGS model learns exclusively from geometrically consistent virtual observations, effectively injecting angular constraints that regularize the view-dependent appearance parameters and suppress the generation of floaters in unobserved regions.

3. Experiment

3.1 Experimental Setting

LLFF Dataset. The LLFF dataset (Mildenhall et al., 2019) comprises eight forward-facing real-world scenes. Following the standard sparse-view evaluation protocol established by FSGS (Zhu et al., 2024), we designate every eighth image as the test set. For training, we evenly sample exactly 3 views from the remaining images. All images are downsampled by a factor of 8, resulting in a resolution of 504×378 . To construct our virtual observations without causing data leakage, we exclude the selected training and testing images, and uniformly sample 5 held-out camera poses from the remaining unused pool in each scene to serve as the virtual camera poses.

LOD-3DGS SmallCampus Dataset. This dataset features a large-scale underground parking garage scenario, providing colorized LiDAR point clouds, perspective images, and COLMAP-format camera parameters at a resolution of 734×734 . To evaluate the model's generalization capability across significant viewpoint variations (e.g., lane changing), we design a spatially shifted evaluation protocol. Specifically, we sample 500 images along the main walking trajectory to form the training set. For the virtual camera poses, we sample 400 images from an adjacent parallel trajectory with a 1-meter lateral offset to the right. Finally, the test set consists of 200 images sampled from a third parallel trajectory, shifted an additional 1 meter to the right (i.e., 2 meters away from the training trajectory). All selected trajectories share the same forward-facing direction.

Evaluation Metrics. To quantitatively assess the quality of novel view synthesis, we employ three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018).

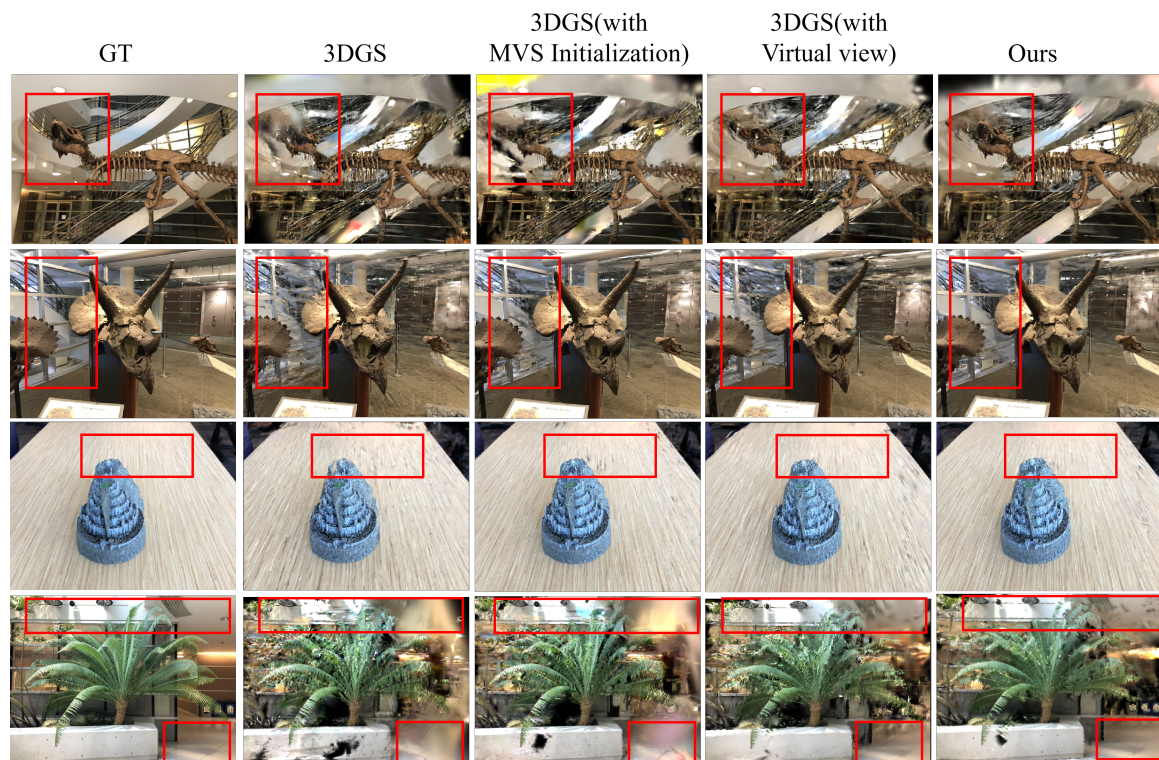


Figure 2. Qualitative results on LLFF Dataset.

Implementation Details. Our method is implemented in PyTorch based on the official 3DGS framework. The initial sparse and dense point clouds are extracted using COLMAP. As described in Sec. 2.3, the proxy 3DGS model for virtual view synthesis is initialized from the dense MVS point cloud, enabling the rendering of color images and binary masks via α -blending. For the main optimization, we train the model for 10,000 iterations on the LLFF dataset and 30,000 iterations on the LOD-3DGS dataset, keeping other hyperparameters at their default settings. All experiments are conducted on a single NVIDIA Tesla T4 GPU.

3.2 Ablation Studies on LLFF Dataset

To validate our core components, we conduct an ablation study on the LLFF dataset under a sparse 3-view setup (Table 1). The baseline 3DGS struggles severely, initializing a mere 167 points and yielding a poor PSNR of 15.284 dB. Introducing MVS initialization resolves this geometric collapse by increasing the initial point cloud density by over $150\times$ (to 26,028 points). This dense structural prior directly translates to a quantitative leap, improving PSNR by 0.912 dB and SSIM by 0.104.

Further integrating Virtual View augmentation (our full VISTA-GS pipeline) yields the most significant performance boost. Compared to the baseline, the full model achieves a significant total gain of +1.670 dB in PSNR (reaching 16.954 dB), +0.120 in SSIM, and a 16.6% reduction in LPIPS error (from 0.421 to 0.351). Notably, the virtual views provide a larger PSNR gain when applied to the MVS initialization (+0.758 dB) than to the baseline (+0.681 dB), indicating a strong positive synergy between the two modules.

Remarkably, despite optimizing a vastly denser point cloud, the full pipeline accelerates convergence, reducing the training time

by 12.6% compared to the baseline (from 3m26s to 3m00s). This proves that virtual views effectively regularize the optimization space, preventing wasteful iterations on flawed geometry. Qualitatively, this synergy successfully eliminates floating artifacts and view-dependent distortions at extreme angles where baseline methods fail completely.

The qualitative comparisons in Figure 2 further illustrate the distinct contribution of each component. The sparse-input baseline produces severe stretching artifacts and incomplete scene structures, particularly in background regions. With MVS dense point cloud initialization, the background completeness improves noticeably; however, as highlighted by the red boxes, pronounced color discontinuities persist across view-point changes—visible, for instance, in the ceiling reflections of the dinosaur skeleton scene (Row 1), the glass display case behind the triceratops skull (Row 2), the wood-grain texture surrounding the blue ornament (Row 3), and the background foliage of the palm tree scene (Row 4). When virtual view augmentation is applied alone, these color jumps are substantially suppressed, and the background coherence is better preserved, confirming that the synthesized views provide effective appearance-level regularization independent of geometric densification.

Our full VISTA-GS pipeline, combining both modules, delivers the best overall visual quality: the foreground subjects (e.g., bone structures, ornament surface details, and palm fronds) are reconstructed with higher fidelity, while the background regions remain structurally complete and color-consistent. Large-area stretching artifacts are effectively eliminated, and the abrupt color transitions observed in the MVS-only variant are thoroughly resolved, demonstrating the complementary synergy between geometric densification and appearance regularization.

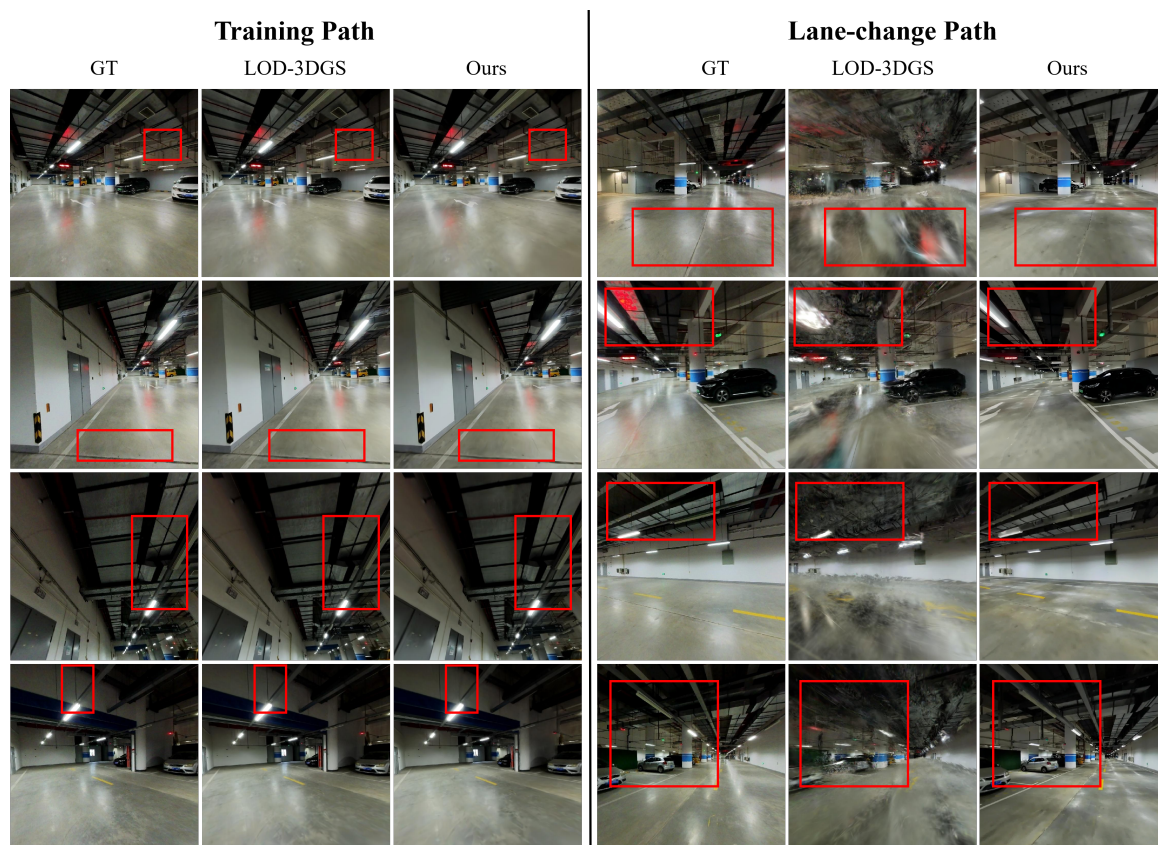


Figure 3. Qualitative results on training path and lane-change path.

Table 1. Quantitative results of ablation study on LLFF Dataset.

MVS Initialization	Virtual View	Points number	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Training Time $\downarrow$
×	×	167	15.284	0.401	0.421	3m26s
×	✓	167	15.965	0.439	0.401	3m06s
✓	×	26028	16.196	0.505	0.353	4m03s
✓	✓	26028	16.954	0.521	0.351	3m

3.3 Comparison on the LiDAR-Initialized Dataset

To evaluate our method’s generalization capabilities in complex real-world environments, we conduct experiments on the LOD-3DGS SmallCampus dataset. As shown in Table 2, baseline methods suffer from severe overfitting to the training trajectory. While incorporating LiDAR initialization improves the Training Path PSNR (from 34.418 dB to 35.073 dB), it actually degrades rendering quality on the unseen Lane-change Path (dropping from 17.477 dB to 17.228 dB). This indicates that relying solely on geometric priors fails to resolve view-dependent appearance overfitting.

To break this overfitting cycle, we synthesize virtual views from colorized LiDAR point clouds. These generated views serve as highly reliable pseudo-labels, achieving an average PSNR of 21.2 dB against the original images. Integrating these virtual views introduces robust data-level constraints that effectively regularize the optimization space. Although this strong regularization deliberately trades off Training Path performance (a 2.491 dB decrease in PSNR), it yields a substantial generalization leap on the unseen Lane-change Path: PSNR surges by +4.489 dB (from 17.228 dB to 21.717 dB), SSIM improves by 20.5% (from 0.593 to 0.715), and LPIPS drops by 21.4% (from 0.480 to 0.377) compared to the LiDAR-only baseline. Finally,

we observe that additionally incorporating depth loss does not yield further improvements, slightly reducing the Lane-change PSNR to 21.692 dB.

The qualitative results presented in Figure 3 further corroborate these quantitative findings. On the Training Path (left three columns), our method achieves rendering quality that is visually comparable to the baseline, with only subtle differences in fine-grained structural details such as ceiling-mounted pipes, thin hangers, and floor lane markings highlighted by the red boxes. This confirms that the virtual-view regularization does not significantly compromise the reconstruction fidelity along the original training trajectory.

In stark contrast, on the Lane-change Path (right three columns), the baseline methods exhibit catastrophic quality degradation. The rendered images are dominated by dense floating artifacts (floaters) that severely occlude the scene content, accompanied by pronounced color bleeding and abrupt color jumps. Structural elements such as ceiling beams, elevator doors, pillars, and parked vehicles become barely recognizable, and the overall scene coherence is almost entirely lost. Our method, shown in the rightmost column, dramatically alleviates these issues: the floaters are substantially suppressed, color consistency is well preserved across viewpoint changes, and critical scene structures—including walls, ceiling light strips, floor markings, and vehicles—are rendered with significantly improved completeness and geometric fidelity. These visual improvements demonstrate that the synthesized virtual views provide effective multi-view constraints that regularize the Gaussian optimization, enabling robust generalization to novel viewpoints far beyond the training trajectory.

Table 2. Quantitative comparison on Lod-3DGS SmallCampus Dataset.

Method	Lidar Initialization	Depth Map	Virtual View	Training Path			Lane-change Path		
				PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	×	×	×	34.418	0.935	0.200	17.477	0.616	0.472
3DGS	✓	×	×	35.073	0.944	0.161	17.228	0.593	0.480
LOD-3DGS	✓	✓	×	34.426	0.946	0.207	17.341	0.575	0.494
Ours	✓	×	✓	32.582	0.930	0.191	21.717	0.715	0.377
Ours (w depth)	✓	✓	✓	32.093	0.925	0.199	21.692	0.714	0.381

Table 3. Comparison of geometric accuracy of different methods.

Method	Lidar Initialization	Depth Map	Virtual View	MAE (cm) ↓	Std. (cm) ↓	Lane-change Path		
						PSNR ↑	SSIM ↑	LPIPS ↓
3DGS	×	×	×	31.82	296.99	17.477	0.616	0.472
3DGS	✓	×	×	9.81	27.17	17.228	0.593	0.480
LOD-3DGS	✓	✓	×	0.44	5.75	17.341	0.575	0.494
Ours	✓	×	✓	9.72	25.42	21.717	0.715	0.377
Ours (w depth)	✓	✓	✓	4.84	18.69	21.692	0.714	0.381

3.4 Comparison of Geometric Accuracy

We evaluate the geometric accuracy of different methods by computing the point-to-point distance between the extracted 3D Gaussian centers and the ground-truth LiDAR point clouds using CloudCompare. The Mean Absolute Error (MAE) and Standard Deviation (Std.) of these distances are reported in Table 3.

Interestingly, our quantitative results reveal a crucial insight into differentiable rendering: higher geometric accuracy does not necessarily translate to better generalization capabilities. For instance, although LOD-3DGS achieves extreme geometric precision with an MAE of merely 0.44 cm, its Lane-change PSNR only reaches 17.341 dB, showing a very limited gain over the LiDAR-initialized 3DGS (17.228 dB). This indicates that simply enforcing stricter geometric constraints fails to significantly improve the image rendering quality for unseen views.

Our proposed virtual view augmentation successfully navigates this dilemma. Compared to the LiDAR-initialized 3DGS, our base model (Ours) slightly improves geometric accuracy (MAE from 9.81 cm to 9.72 cm) while unlocking a substantial +4.489 dB surge in Lane-change PSNR. This demonstrates that our pseudo-GT virtual views provide just enough regularization to prevent structural collapse without overly constraining the differentiable rendering process.

Furthermore, we explore the explicit trade-off between geometry and rendering by introducing an additional depth loss (Ours w/ depth). As shown in Table 3, this depth supervision drastically cuts the geometric error by half, reducing the MAE by 50.2% (from 9.72 cm to 4.84 cm) and the Std. by 26.5% (from 25.42 cm to 18.69 cm). Remarkably, this substantial leap in geometric fidelity comes at a negligible cost of only 0.025 dB in Lane-change PSNR (dropping from 21.717 dB to 21.692 dB). Ultimately, by initializing from LiDAR priors and combining virtual views with depth constraints, our method strikes an optimal balance—maintaining highly competitive geometric accuracy while significantly enhancing rendering robustness on unseen trajectories.

4. Conclusion

In this paper, we presented VISTA-GS, a novel framework designed to tackle the severe overfitting and artifact issues inher-

ent in 3D Gaussian Splatting under sparse-view and restricted-trajectory conditions (e.g., training data strictly limited to a single walking trajectory). Our primary objective is to significantly enhance rendering robustness and image quality for out-of-distribution (OOD) novel views that deviate substantially from the training set. By synergizing Multi-View Stereo (MVS) initialization with a physically-grounded virtual view augmentation strategy, our method effectively introduces crucial angular constraints into the underconstrained optimization process. Furthermore, the use of masked photometric supervision ensures that the model learns from geometrically consistent regions without being misled by synthesis artifacts. Extensive experiments demonstrate the superiority of VISTA-GS.

Notably, Our analysis reveals two important insights. First, higher geometric accuracy does not necessarily translate to better generalization under large viewpoint deviations; appearance overfitting on the limited training views can dominate the rendering quality at novel angles, regardless of how precise the underlying geometry is. Second, even when virtual view augmentation is applied as a standalone component, the color discontinuities commonly observed in OOD renderings are substantially suppressed, and the overall rendering completeness and color consistency are well preserved. These findings confirm that explicitly constraining view-dependent appearance optimization is at least as critical as improving geometry for achieving robust novel-view synthesis in sparse, single-trajectory scanning scenarios.

For future work, we plan to integrate vision foundation models to improve the completeness of MVS-generated dense point clouds, particularly in texture-less and background regions. Additionally, noting that the PSNR of our current synthesized virtual views is limited to approximately 21.2 dB, we aim to leverage advanced generative models to optimize these views, generating more photorealistic images to ensure stricter multi-view consistency. Finally, to further mitigate overfitting, we will explore more rational and adaptive strategies for generating virtual camera poses, thereby improving rendering robustness under larger viewpoint deviations.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Project Nos. 42471442, 42171265,

42371431), Research Project of Shenzhen S and T Innovation Committee (Project No. KJZD20230923115508017), Natural Science Foundation of Guangdong Province (Project No. 2024A1515030061), Project of China State Railway Group Co., Ltd. Science and Technology Research and Development (No. N2024S008), and Research project of State Key Laboratory of Subtropical Building and Urban Science (Project No. 2023ZB18)

References

- Cui, J., Cao, J., Zhao, F., He, Z., Chen, Y., Zhong, Y., Xu, L., Shi, Y., Zhang, Y., Yu, J., 2024. LetsGo: Large-Scale Garage Modeling and Rendering via LiDAR-Assisted Gaussian Primitives. *ACM Trans. Graph.*, 43(6).
- Fei, T., Bi, L., Gao, J., Xin, Q., Zhang, G., 2025. MVSGS: Gaussian Splatting Radiation Field Enhancement Using Multi-View Stereo. *Complex & Intelligent Systems*, 11, 80.
- He, Z., Xiao, Z., Chan, K.-C., Zuo, Y., Xiao, J., Lam, K.-M., 2025. See in detail: Enhancing sparse-view 3d gaussian splatting with local depth and semantic regularization. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1–5.
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).
- Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L., 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20775–20785.
- Mildenhall, B., Srinivasan, P. P., Ortiz-Cayon, R., Kalantari, N. K., Ramamoorthi, R., Ng, R., Kar, A., 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4), 1–14.
- Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X. et al., 2025. Recondreamer: Crafting world models for driving scene reconstruction via on-line restoration. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1559–1569.
- Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Song, M., Lin, X., Zhang, D., Li, H., Li, X., Du, B., Qi, L., 2025. D²GS: Depth-and-Density Guided Gaussian Splatting for Stable and Accurate Sparse-View Reconstruction. *arXiv preprint arXiv:2510.08566*.
- Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.
- Wu, J. Z., Zhang, Y., Turki, H. et al., 2025. DiFix3D+: Improving 3D reconstructions with single-step diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26024–26035.
- Xu, W., Gao, H., Shen, S., Peng, R., Jiao, J., Wang, R., 2024. Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. *European Conference on Computer Vision*, Springer, 203–220.
- Yu, H., Long, X., Tan, P., 2024. Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. *arXiv preprint arXiv:2409.03456*.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhu, Z., Fan, Z., Jiang, Y., Wang, Z., 2024. Fsgs: Real-time few-shot view synthesis using gaussian splatting. *European conference on computer vision*, Springer, 145–163.