StereoGS: Sparse-View 3D Gaussian Splatting via Stereo Priors

Wenhao Yuan, Yiyuan Ge, and Deli Cai*

South China University of Technology, China
wenhaoyuan.stringer@gmail.com, yiyuange@mail.scut.edu.cn,
eecaideli@mail.scut.edu.cn

Abstract. 3D Gaussian Splatting (3DGS) has achieved remarkable success in real-time novel view synthesis, yet it suffers from severe overfitting under sparse-view settings due to insufficient geometric constraints. While recent methods introduce monocular depth priors to mitigate this, they inherently struggle with scale ambiguity and cross-view inconsistency, leading to defective geometry. In this paper, we propose StereoGS, a novel sparse-view 3DGS framework that integrates stereo priors to establish reliable binocular consistency. Unlike scale-agnostic monocular constraints, StereoGS introduces a Stereo Depth Regularization by constructing virtual stereo pairs during optimization and leveraging a foundation stereo model to enforce absolute scale and binocular-consistent structures. To further suppress overfitting and eliminate redundant primitives, we design a Gradient-Aware Opacity Decay strategy that dynamically penalizes Gaussians based on their relative opacity gradient magnitudes. Combined with a Consistency-Aware Dense Initialization using zero-shot multi-view depth estimation, StereoGS effectively anchors primitives to accurate scene surfaces. Extensive experiments on LLFF, DTU, Mip-NeRF360, and Blender datasets demonstrate that StereoGS achieves state-of-the-art performance in sparse-view settings without incurring any additional inference overhead. Project Page: https://stringerywh00.github.io/StereoGS_project_page/

Keywords: Novel View Synthesis · 3D Gaussian Splatting · Sparse View Synthesis

1 Introduction

Novel View Synthesis (NVS) [1, 23, 35] aims to synthesize realistic images from unseen viewpoints. While Neural Radiance Field (NeRF) [19] and 3D Gaussian Splatting (3DGS) [13] have revolutionized high-fidelity NVS, 3DGS has gained particular attention for its real-time rendering capabilities. However, vanilla 3DGS relies heavily on dense input views. Under sparse-view settings [7, 23, 35, 50], limited view overlap and insufficient appearance correspondences restrict sufficient geometric constraints, inevitably leading to severe overfitting.

* Corresponding author.

Consequently, achieving robust sparse-view 3DGS remains a critical yet challenging task for real-world applications.

To mitigate this, existing methods for sparse-view 3DGS [4, 17, 50] attempt to introduce geometric priors by using monocular depth estimation for geometry regularization. While effective, we argue that relying solely on monocular cues suffers from two fundamental limitations. First, monocular depth inherently suffers from scale ambiguity, producing relative rather than absolute depth estimates. Since these relative depth maps lack a unified metric scale, enforcing them as regularization often conflicts with photometric consistency derived from the input images, thereby misguiding the optimization and leading to defective geometry. Second, monocular depth maps lack cross-view consistency because they are inferred independently for each view. Consequently, the same spatial point is frequently assigned inconsistent depth values across different viewpoints. When back-projected into 3D space, these view-wise inconsistencies lead to spatial misalignments, destabilizing the Gaussian optimization and causing severe artifacts.

Recent studies [7,39,48] attempt to integrate cross-view consistency for Gaussian optimization. For instance, Binocular3DGS [7] applies photometric loss on binocular images via the rendered depth map. However, this indirect pixel-level constraint lacks direct gradient guidance to 3D Gaussian primitives, failing to anchor them accurately to scene surfaces. NexusGS [48] leverages the traditional epipolar geometry for initialization, but relying solely on initialization without continuous binocular consistency constraints during optimization still results in geometric degradation, especially in textureless or occluded regions.

In this paper, we propose **StereoGS**, a sparse-view Gaussian Splatting framework that integrates stereo priors to establish reliable binocular consistency during Gaussian optimization. StereoGS is optimized via stereo priors during both initialization and optimization, compelling the 3D Gaussians to learn scale-accurate geometry. Specifically, unlike previous methods that rely on scale-agnostic monocular depth constraints, StereoGS regularizes the geometry by imposing a novel stereo depth constraint during optimization. For each training view, we construct a virtual right camera to form a stereo pair and render the right-view image, and depth maps for training views. By feeding the RGB pair of the ground-truth left-view and the rendered right-view into a foundation stereo model [37], we obtain a highly accurate reference depth map that serves as an external stereo depth prior. We then perform supervision between the rendered and reference depth maps. This process effectively integrates a scale-aware stereo depth prior into the optimization, compelling the 3D Gaussians to learn absolute scale and binocular-consistent structures.

To suppress overfitting and eliminate redundant primitives, we introduce a general gradient-aware opacity decay strategy. Our core insight is that the opacity gradient of a Gaussian inherently reflects its necessity for scene reconstruction. Therefore, we dynamically penalize Gaussians based on their gradient magnitudes: primitives with smaller gradients are deemed less critical for rendering and receive a stronger decay penalty, while those with larger update demands

are preserved. These mechanisms, alongside the stereo priors, serve as structural signals that backpropagate through the rendering pipeline to Gaussian primitives. As a result, StereoGS learns highly accurate, binocular-consistent geometric representations, directly mitigating the severe overfitting inherent in sparse-view settings.

Additionally, for Gaussian initialization, we employ a pre-trained multi-view depth estimator [9] combined with strict consistency filtering to estimate the multi-view-consistent depth maps for each view. By further applying cross-view reprojection errors, we eliminate outliers in the depth maps and fuse them into a dense point cloud, which serves as a robust initialization.

Notably, because the approaches in StereoGS are training-time strategies, employing StereoGS incurs no additional computational overhead during inference compared to the vanilla 3DGS. Extensive experiments on LLFF, DTU, Mip-NeRF360, and Blender datasets demonstrate that StereoGS achieves state-of-the-art performance in sparse-view settings, effectively balancing geometric accuracy with rendering quality.

Our main contributions are summarized as follows:

- We introduce a Stereo Depth Regularization that utilizes stereo priors to enforce absolute scale constraints during optimization, fundamentally resolving the scale ambiguity inherent in monocular depth priors.
- We design Gradient-Aware Opacity Decay to dynamically adjust opacity during training based on the gradient, effectively eliminating redundant primitives and suppressing overfitting to ensure a compact and robust scene representation.
- Extensive experiments on LLFF, DTU, Mip-NeRF360, and Blender datasets demonstrate that our method achieves state-of-the-art results compared to existing sparse-view methods.

2 Related Works

2.1 Radiance Fields

Radiance Fields are employed for reconstructing 3D scenes and synthesizing novel views. Neural Radiance Fields (NeRFs) [19] have seen significant advances that learn neural volumetric representations of 3D scenes and render images via volume rendering. While NeRF enables high-quality reconstruction from images, it requires high computational costs and suffers from slow training and inference speeds. Subsequent research has focused on addressing these bottlenecks by improving the rendering quality [1, 2], computational efficiency [21, 33], frequency regularization [41].

A significant breakthrough in achieving real-time rendering is 3D Gaussian Splatting (3DGS) [13], which represents scenes using a set of explicit 3D Gaussian primitives. 3DGS reconstructs high-quality scenes rapidly by utilizing differentiable splatting, excelling particularly in handling high-frequency details and offering intuitive interpretability. Building on its significant performance,

many follow-up works focus on improving the rendering details [14, 40, 44, 47], compressing the Gaussians and accelerating rendering [16, 22].

2.2 Novel View Synthesis from Sparse View

NeRF-based and 3DGS-based methods have demonstrated impressive rendering performance, whereas in real-world scenarios, only a small number of images are available (i.e., sparse input views). Directly applying the methods in a dense-view setting leads to performance degradation due to overfitting. Early studies have investigated the regularization of NeRFs under sparse-view conditions [5, 15, 23, 41]. Other approaches [38, 49] leverage generative models that are pre-trained on large-scale datasets to distill generative prior into 3D representation. In contrast, alternative strategies [5,28,31,35] leverage external geometric priors derived from pre-trained models (e.g., depth priors) to guide and regularize the training process.

Recently, 3D Gaussian Splatting has been adapted to sparse settings, with a primary focus on initialization, depth regularization, and structural constraints. For Gaussian initialization, Chung *et al.* [4] and CoherentGS [24] estimate monocular depth maps and refine them to achieve more accurate initialization. DNGaussian [17] and FSGS [50] apply depth regularization based on monocular depth estimators. However, relying on monocular depth often lacks absolute scale depth perception. Although MVPGS [39] leverage Multi-View Stereo (MVS) depth estimation to obtain a well-initialized, view-consistent point cloud, the absence of continuous geometric regularization during optimization causes this high-quality initial structure to gradually deteriorate. Other approaches, such as Gaussian pruning based on point disagreement [45] and a random dropout strategy [25], also effectively mitigate overfitting to known yet sparse views. To integrate cross-view consistency during optimization, Binocular3DGS [7] warps rendered RGB images and applies a photometric loss between the constructed binocular RGB images. We argue that this self-supervision constraint on RGB images lacks explicit geometric constraints and leads to geometric artifacts. Despite improvements, existing methods predominantly lack explicit cross-view consistency constraints during Gaussian optimization and can lead to inevitable geometric artifacts. In contrast, we propose integrating a stereo constraint from the stereo matching prior into the optimization process.

3 Method

3.1 Preliminary for 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [13] models the 3D scene as a collection of anisotropic 3D Gaussians, serving as an explicit scene representation. Each primitive G_i is spatially defined by a center position μ and a 3D covariance matrix Σ. The influence of a Gaussian at a spatial coordinate x is formulated as:

$$G(x) = \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{1}$$

To ensure the positive semi-definite property of the covariance matrix during optimization, Σ is factorized into a scaling matrix S and a rotation matrix R, such that $\Sigma = RSS^\top R^\top$. Additionally, each Gaussian encodes view-dependent appearance via spherical harmonic (SH) coefficients and a learnable opacity α.

To synthesize novel views, these 3D Gaussians are projected into the 2D image plane via differentiable splatting. Given a viewing transformation W and the Jacobian of the affine approximation of the projective transformation J, the 3D covariance Σ is transformed into a 2D planar covariance $\Sigma' = JW\Sigma W^\top J^\top$. The rendering pipeline employs a tile-based rasterizer that sorts the Gaussians overlapping a specific pixel by depth. The final pixel color C is derived through α-blending, accumulating the contribution of N ordered Gaussians:

$$C = \sum_{i=1}^{N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{2}$$

where c_i denotes the color decoded from SH coefficients, and α'_i represents the effective 2D opacity.

Similarly, the final pixel depth D can be rendered by accumulating the projected depth values d_i (the distance from the camera center to the Gaussian center) using the same blending weights:

$$D = \sum_{i=1}^{N} d_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j). \tag{3}$$

3D Gaussian Splatting is typically trained end-to-end by minimizing the discrepancy between the rendered and ground-truth images using a combination of $\mathcal{L}_1$ and D-SSIM loss terms.

3.2 Stereo Depth Regularization

Geometric inaccuracy remains a critical bottleneck in sparse-view 3D Gaussian Splatting. Several prior works [17, 50] incorporate monocular depth [26, 27, 42] as a regularization term during the optimization process. However, the reconstructed geometry often remains suboptimal due to the inherent scale ambiguity and cross-view inconsistency of these monocular priors. To address this limitation, recent approaches [7] enforce stereo consistency using photometric constraints. They warp synthesized views to the input views using the rendered depth to minimize reconstruction errors. However, this warping-based approach lacks direct supervision on Gaussian geometry and is prone to failure in textureless or repetitive regions.

To overcome these limitations, we propose stereo depth regularization. By constructing virtual stereo pairs and leveraging a foundation stereo model [37], we derive reliable stereo depth to regularize the Gaussian geometry, as illustrated in Fig. 1 (a).

Stereo Camera Pair Construction. During optimization, we treat the training view as the left camera and synthesize a virtual right camera by applying a

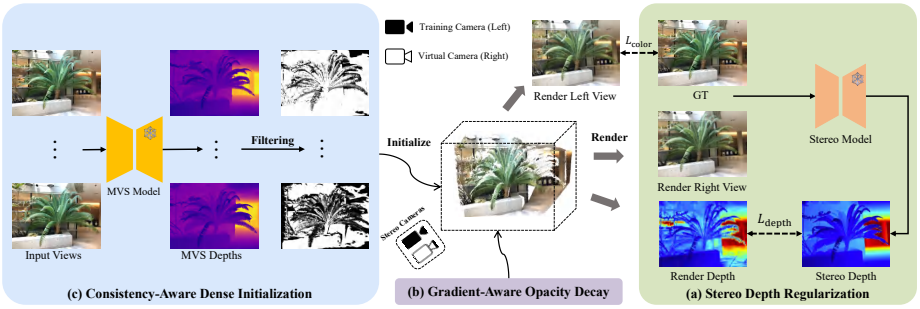

Fig. 1: Overview of StereoGS, comprising three components: (a) Stereo Depth Regularization synthesizes virtual stereo pairs to extract stereo depth, enforcing explicit geometric consistency between paired views during optimization. (b) Gradient-Aware Opacity Decay dynamically penalizes Gaussian opacities according to their gradients, facilitating the removal of redundant primitives and mitigating overfitting. (c) Consistency-Aware Dense Initialization leverages multi-view depth priors and geometric filtering to construct a dense and reliable geometric foundation.

horizontal translation. We then render the corresponding right-view image $\hat{I}_r$ using the current 3D Gaussians, forming a stereo pair $(I_l, \hat{I}_r)$ with the ground-truth left image I_l.

Left-Right Consistency Check. We employ FoundationStereo [37] as our pre-trained stereo matching model Φ. The left-view disparity is computed as $\hat{D}_l = \Phi(I_l, \hat{I}_r)$. We explicitly use the clean ground-truth left image instead of a rendered one because the estimated disparity serves as an absolute depth reference; avoiding optimization noise and rendering artifacts ensures a highly reliable geometric prior.

To filter out unreliable predictions, we apply the left-right consistency check that is widely applied in stereo vision literature [6, 8, 11]. We simultaneously extract the right-view disparity. Since standard stereo models typically output left-view disparity, we introduce a horizontal flipping operator $\mathcal{F}(\cdot)$ to compute it: $\hat{D}_r = \mathcal{F}\left(\Phi\left(\mathcal{F}(I_l), \mathcal{F}(\hat{I}_r)\right)\right)$. The left-right consistency effectively identifies occluded regions and eliminates false matches by judging whether the disparity discrepancy between $\hat{D}_l$ and $\hat{D}_r$ exceeds a threshold, forming an occlusion mask M_{occ}. For specific details, please refer to the supplementary material.

To further refine the valid regions, we compute a background mask M_{bg} and a disparity anomaly mask M_{anomaly}. The final validity mask for depth regularization is derived by fusing these masks: $M_{\text{valid}} = 1 - (M_{\text{bg}} \vee M_{\text{occ}} \vee M_{\text{anomaly}})$.

Depth Regularization. We convert the estimated disparity $\hat{D}_l$ into depth Z_{stereo} using the standard triangulation formula: $Z_{\text{stereo}} = fd/\hat{D}_l$, where f is the focal length and d is the baseline between camera pairs. This Z_{stereo} serves as pseudo-ground-truth and the stereo prior from the external stereo model. To further improve numerical stability during loss computation and to better

constrain the geometric accuracy in near-foreground regions, we minimize the L_1 error in the inverse depth space between Z_{stereo} and the rendered left-view depth $\hat{Z}$ for all valid pixels, using the validity mask M_{valid}.

$$\mathcal{L}_{\text{depth}} = \| M_{\text{valid}} \odot (\frac{1}{\hat{Z}} - \frac{1}{Z_{\text{stereo}}}) \|_1, \tag{4}$$

3.3 Gradient-Aware Opacity Decay

Opacity determines the visibility and contribution of each Gaussian primitive to the rendered image. Despite the importance of this attribute, existing methods [4,13,17,39,48,50] predominantly rely on a periodic opacity reset mechanism to eliminate noisy primitives. While effective in dense-view scenarios, this aggressive reset indiscriminately suppresses both stable surface structures and transient floaters in sparse-view settings. Although recent sparse-view techniques [7] employ a strategy with fixed decay rate that penalizes all Gaussians equally, this approach cannot effectively evaluate the importance of individual Gaussians or distinguish useful geometry from noise.

Motivated by these limitations, we propose a gradient-aware opacity decay strategy. Our core insight is that the opacity gradient magnitude indicates a Gaussian primitive's importance: Gaussians with large opacity gradients contribute significantly to reducing photometric error and should be preserved, whereas those with negligible gradients likely represent redundant floaters or noise.

To implement this insight, we derive a dynamic decay factor γ by incorporating the local gradient magnitude into a predefined base factor γ_{base}. This factor γ is then used to modulate the original opacity α. Formally, let $g = |\nabla_\alpha \mathcal{L}|$ denote the magnitude of the gradient of the loss function $\mathcal{L}$ with respect to the opacity α. We observe that the absolute values of g are typically minuscule (e.g., on the order of 10^{-6}). To ensure robustness against such scale variations, we guide the opacity decay using a relative gradient magnitude. Inspired by Group Relative Policy Optimization (GRPO) [30], which evaluates the output actions based on their relative advantages within a group rather than relying on an absolute value network, we compute a relative gradient $\beta = g/\bar{g}$ for each Gaussian, where $\bar{g}$ is the mean opacity gradient magnitude across all Gaussians at the current iteration. Incorporating a basic decay factor γ_{base}, we derive a dynamic decay factor γ via an exponential soft-thresholding function:

$$\gamma = 1 - (1 - \gamma_{\text{base}}) \exp\left(-s \cdot \beta\right), \tag{5}$$

where s is a sensitivity hyperparameter controlling the decay curvature. Finally, the opacity is scaled by $\hat{\alpha} = \gamma\alpha$.

3.4 Consistency-Aware Dense Initialization

Previous methods [4,13] typically utilize sparse point clouds generated by Structure-from-Motion (SfM) [29] for initialization. However, in sparse-view settings, SfM

often yields point clouds that are too sparse and noisy to adequately represent the scene. While some recent works [7,39] have explored using MVS priors (e.g., MVS models [3] and keypoint matching networks [34]) to improve initialization, their performance is often constrained by the domain gap between pre-training data and diverse in-the-wild scenes, as illustrated in Fig. 1 (c).

Building on this MVS-based paradigm, we overcome the generalization bottleneck by leveraging a more advanced zero-shot model, MVSAnywhere [9], to extract multi-view-consistent depth maps. Specifically, for each training view (treated as the target image), we input it along with the remaining source images into the MVS model to estimate its depth map. By iterating over all training views, we obtain a set of cross-view consistent depth maps. We then adopt a geometric filtering strategy [43] based on cross-view reprojection errors to eliminate outliers. The filtered depth maps are finally back-projected and fused into a unified dense point cloud. Thanks to the robust zero-shot capabilities of MVSAnywhere, we obtain higher-quality point clouds compared to [39] and [7], providing a significantly more robust initialization for the 3D Gaussians. We provide visual comparisons of the initialization point clouds in the experiment section.

3.5 Training Loss

The total objective function comprises two components: the proposed stereo matching prior-guided depth regularization $\mathcal{L}_{\text{depth}}$, as detailed in Eq. (4), and the color reconstruction loss $\mathcal{L}_{\text{color}}$ adopted from 3DGS [13]. We define the overall loss $\mathcal{L}$ as:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{depth}}, \tag{6}$$

where $\mathcal{L}_{\text{color}}$ represents a linear combination of the $\mathcal{L}_1$ loss and the D-SSIM loss $\mathcal{L}_{\text{D-SSIM}}$, formulated as:

$$\mathcal{L}_{\text{color}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}. \tag{7}$$

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on four public datasets: LLFF [18], DTU [10], Mip-NeRF360 [2], and Blender [19]. Following prior works [7, 17, 23, 50], we used 3, 6, and 9 views as training sets for the LLFF and DTU datasets, 12 and 24 views for the Mip-NeRF360 dataset, and 8 images for training on the Blender dataset. The selection of test images remained consistent with previous works [7,17,23,50]. The downsampling rates for the LLFF, DTU, Mip-NeRF360, and Blender datasets are 8, 4, 8, and 2, respectively.

Baselines. We compare the proposed StereoGS with representative existing methods, including NeRF-based approaches (RegNeRF [23], FreeNeRF [41], and SparseNeRF [35]) and 3DGS-based approaches (3DGS [13], DNGaussian [17],

Table 1: Quantitative comparison on LLFF with 3, 6, 9 training views. We mark the best, second best, and third best methods in cells, respectively.

Methods	3-view			6-view			9-view		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RegNeRF [23]	19.08	0.587	0.336	23.10	0.760	0.206	24.86	0.820	0.161
FreeNeRF [41]	19.63	0.612	0.308	23.73	0.779	0.195	25.13	0.827	0.160
SparseNeRF [35]	19.86	0.624	0.328	23.26	0.741	0.235	24.27	0.781	0.228
3DGS [13]	16.02	0.465	0.378	19.45	0.627	0.268	21.13	0.715	0.214
DNGaussian [17]	19.12	0.591	0.294	22.18	0.755	0.198	23.17	0.788	0.180
FSGS [50]	20.31	0.652	0.288	24.20	0.811	0.173	25.32	0.856	0.136
CoR-GS [45]	20.45	0.712	0.196	24.49	0.837	0.115	26.06	0.874	0.089
MVPGS [39]	20.54	0.727	0.194	23.64	0.819	0.137	24.23	0.843	0.120
DropGaussian [25]	20.76	0.713	0.200	24.74	0.837	0.117	26.21	0.874	0.088
NexusGS [48]	21.07	0.738	0.177	-	-	-	-	-	-
Binocular3DGS [7]	21.44	0.751	0.168	24.87	0.845	0.106	26.17	0.877	0.090
D^2GS [32]	21.35	0.746	0.179	24.84	0.834	0.122	-	-	-
Ours	21.91	0.773	0.157	24.92	0.853	0.104	26.25	0.879	0.087
Ours*	22.05	0.783	0.147	25.40	0.856	0.102	26.44	0.883	0.082

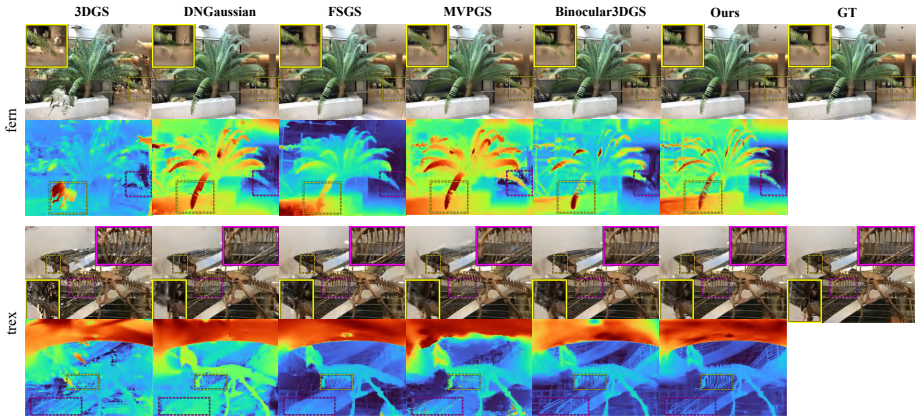

Fig. 2: Visual comparison on LLFF dataset, including RGB images and colored depth maps. The Gaussians are learned under 1/8 resolution with 3 input views.

FSGS [50], CoR-GS [45], MVPGS [39], NexusGS [48], DropGaussian [25], Binocular3DGS [7], and D^2GS [32]). For most evaluated methods, we directly report the best quantitative results from the respective original publications. For the vanilla 3DGS, we report the results obtained from our own implementation. We employ the average values of PSNR, SSIM [36], and LPIPS [46] as the quantitative evaluation metrics.

4.2 Comparison

Comparison on LLFF. Table 1 presents the quantitative results on the LLFF dataset under configurations of 3, 6, and 9 input views. Our StereoGS con-

Table 2: Quantitative comparison on DTU with 3, 6, 9 training views. We mark the best, second best, and third best methods in cells, respectively.

Methods	3-view			6-view			9-view		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RegNeRF [23]	18.89	0.745	0.190	22.20	0.841	0.117	24.93	0.884	0.089
FreeNeRF [41]	19.52	0.787	0.173	23.25	0.844	0.131	25.38	0.888	0.102
SparseNeRF [35]	19.47	0.829	0.183	23.26	0.843	0.135	25.13	0.871	0.114
3DGS [13]	10.99	0.585	0.313	20.33	0.776	0.223	22.90	0.816	0.173
DNGaussian [17]	18.91	0.790	0.176	22.10	0.851	0.148	23.94	0.887	0.131
FSGS [50]	17.34	0.818	0.169	21.55	0.880	0.127	24.33	0.911	0.106
CoR-GS [45]	19.21	0.853	0.119	24.51	0.917	0.068	27.18	0.947	0.045
MVPGS [39]	20.65	0.877	0.099	23.98	0.921	0.066	26.45	0.946	0.047
NexusGS [48]	20.21	0.869	0.102	-	-	-	-	-	-
Binocular3DGS [7]	20.71	0.862	0.111	24.31	0.917	0.073	26.70	0.947	0.052
Ours	21.46	0.879	0.099	24.86	0.926	0.064	26.83	0.955	0.049
Ours*	22.00	0.890	0.095	25.41	0.938	0.063	27.39	0.955	0.047

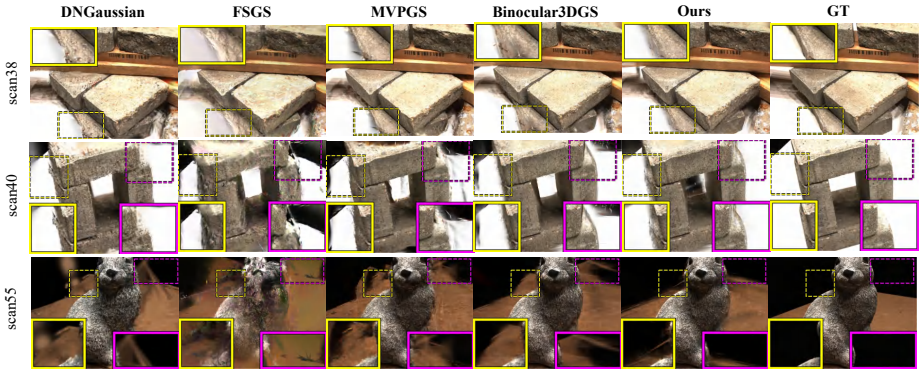

Fig. 3: Visual comparison on DTU under 1/8 resolution with 3 views.

sistently outperforms prior approaches across all view configurations in terms of PSNR, SSIM, and LPIPS. For NexusGS, we could not reproduce their results and report only the 3-view result from their paper. Fig. 2 provides visual comparisons of novel view synthesis and depth rendering for the *fern* and *trex* scenes from the LLFF dataset. As demonstrated by the rendered RGB images, vanilla 3DGS [13] captures only the coarse scene layout and suffers from severe patchy artifacts. For methods that employ monocular depth regularization, such as DNGaussian [17] and FSGS [50], the primary limitation lies in an inability to accurately reconstruct the underlying scene geometry. Although MVPGS [39] and Binocular3DGS [7] utilize MVS to establish an improved geometric initialization, these approaches still struggle during the optimization process. Specifically, Binocular3DGS relies on a self-supervised photometric consistency loss that is inherently unreliable in textureless regions. Furthermore, the use of a fixed-ratio opacity decay rate in Binocular3DGS fails to preserve essential surface Gaussians, as shown by the skeleton of the *trex*. These limitations result

Table 3: Quantitative results on Mip-NeRF360 with 12 and 24 views.

Methods	12-view			24-view		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RegNeRF [23]	18.55	0.524	0.426	22.19	0.643	0.335
FreeNeRF [41]	18.68	0.528	0.421	22.78	0.689	0.323
SparseNeRF [35]	18.73	0.531	0.419	22.85	0.693	0.315
3DGS [13]	18.52	0.523	0.415	22.80	0.708	0.276
FSGS [50]	18.80	0.531	0.418	23.28	0.715	0.274
CoR-GS [45]	19.52	0.558	0.418	23.39	0.727	0.271
NexusGS [48]	-	-	-	23.86	0.753	0.206
DropGaussian [25]	19.74	0.577	0.364	24.05	0.761	0.226
D^2GS [32]	20.09	0.587	0.356	24.13	0.763	0.221
Ours	20.25	0.618	0.312	24.18	0.775	0.222
Ours*	20.51	0.625	0.309	24.25	0.783	0.206

in geometric distortion and subsequent rendering artifacts, ultimately degrading the reconstructed details. In contrast, our StereoGS utilizes stereo depth regularization and employs a gradient-aware opacity decay strategy, effectively resolving ambiguities in textureless regions and preserving critical scene geometry.

Comparison on DTU. Table 2 presents the quantitative results on the DTU dataset with 3, 6, and 9 input views. Our StereoGS achieves state-of-the-art performance across most evaluation metrics. Specifically, under the highly constrained 3-view setting, StereoGS demonstrates significant improvements in PSNR, SSIM, and LPIPS compared to recent approaches such as Binocular3DGS [7] and MVPGS [39]. Fig. 3 provides visual comparisons for three scenes from the DTU dataset. As observed in the rendered novel views, DNGaussian [17], FSGS [50], and MVPGS [39] yield blurry results, whereas Binocular3DGS [7] exhibits numerous artifacts in textureless regions due to self-supervised photometric constraints. In contrast, our StereoGS significantly reduces artifacts in textureless regions while achieving superior rendering quality.

Comparison on Mip-NeRF360. Table 3 presents the quantitative results on the Mip-NeRF360 dataset using 12 and 24 input views. The proposed method achieves superior performance across most evaluation metrics. Specifically, when compared to recent state-of-the-art approaches, such as D^2GS [32] and DropGaussian [25], StereoGS demonstrates significant improvements under both the 12-view and 24-view configurations. Figure 3 provides visual comparisons of novel view synthesis and depth rendering for the *bicycle* and *stump* scenes from the Mip-NeRF360 dataset. In the *bicycle* scene, baseline methods including FSGS [50], CoR-GS [45], and DropGaussian [25] exhibit prominent artifacts in large textureless regions (for example, the grass areas indicated by dashed boxes), whereas the proposed approach maintains high rendering fidelity. Furthermore, in the *stump* scene, these baseline approaches show difficulty in reconstructing fine-grained object details, which are successfully preserved by the proposed method. Consistent with the results of RGB rendering, the proposed method produces significantly more accurate and coherent depth rendering than the baselines.

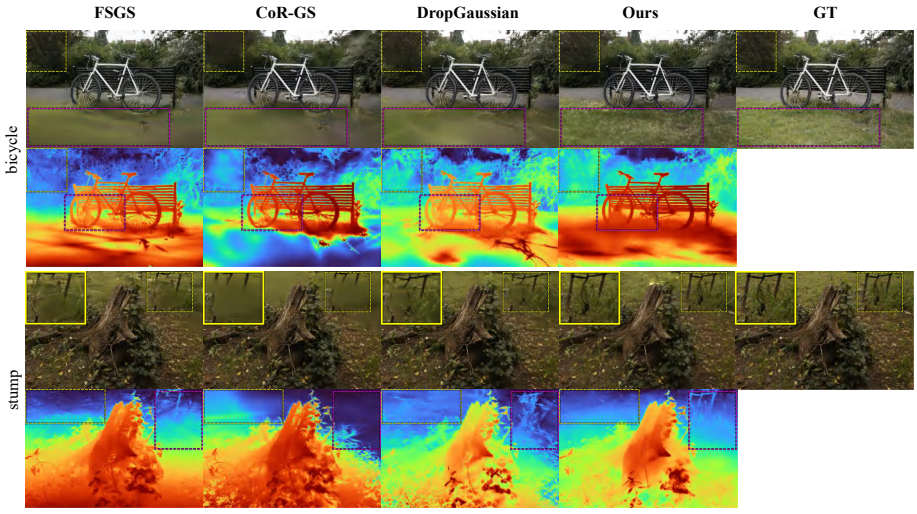

Fig. 4: Visual comparison on the Mip-NeRF360 dataset.

Table 4: Quantitative results on Blender with 8 training views.

Methods	PSNR↑	SSIM↑	LPIPS↓
RegNeRF [23]	23.86	0.852	0.105
FreeNeRF [41]	24.26	0.883	0.098
SparseNeRF [35]	24.04	0.876	0.113
3DGS [13]	23.20	0.870	0.104
DNGaussian [17]	24.31	0.886	0.088
FSGS [50]	24.64	0.895	0.095
CoR-GS [45]	24.43	0.896	0.084
NexusGS [48]	24.37	0.893	0.087
DropGaussian [25]	24.42	0.888	0.089
Binocular3DGS [7]	24.71	0.872	0.101
Ours	24.83	0.899	0.081
Ours*	25.04	0.899	0.078

Comparison on Blender. Table 4 presents the quantitative results on the Blender dataset using 8 input views. Our proposed method consistently outperforms all baselines across all metrics, demonstrating its superior capability in sparse-view reconstruction. Fig. 5 provides visual comparisons for two scenes from the Blender dataset.

4.3 Ablation Studies

To verify the effectiveness of the proposed consistency-aware dense initialization, stereo depth regularization, and gradient-aware opacity decay strategy, we conduct ablation studies by incrementally integrating these components on the LLFF and DTU datasets. As demonstrated in Table 5, the removal of any proposed module leads to a clear degradation in performance.

Fig. 5: Visualization on Blender under 1/2 resolution with 8 input views.

Table 5: Ablation studies on LLFF and DTU datasets with 3 input views.

Consistency-Aware Dense Initialization	Stereo Depth Regularization	Gradient-Aware Opacity Decay	LLFF PSNR↑	LLFF SSIM↑	LLFF LPIPS↓	DTU PSNR↑	DTU SSIM↑	DTU LPIPS↓
✗	✗	✗	16.02	0.465	0.378	10.99	0.585	0.313
✓	✗	✗	19.75	0.691	0.215	14.10	0.786	0.196
✗	✓	✗	17.32	0.524	0.317	12.46	0.697	0.208
✗	✗	✓	18.18	0.569	0.291	15.05	0.751	0.202
✗	✓	✓	18.96	0.605	0.262	17.66	0.784	0.172
✓	✓	✗	19.79	0.695	0.214	15.57	0.812	0.166
✓	✗	✓	21.18	0.741	0.171	19.76	0.863	0.112
✓	✓	✓	21.91	0.773	0.157	21.46	0.879	0.099

Effectiveness of Consistency-Aware Dense Initialization. As shown in the second row of Table 5, incorporating consistency-aware dense initialization substantially improves performance over the baseline (*e.g.*, PSNR increases from 16.02 to 19.75 on the LLFF dataset and from 10.99 to 14.10 on the DTU dataset). As illustrated in Fig. 6, the baseline 3DGS fails to reconstruct meaningful geometry, resulting in disordered depth maps and severe floaters. In contrast, the proposed dense initialization enables the clear recovery of the overall scene structure. This demonstrates that zero-shot MVS priors provide a more robust and crucial geometric foundation than sparse SfM points. Furthermore, to validate specific design choices, we conduct additional ablation studies using different MVS models and compare the generated point clouds with recent state-of-the-art methods in the supplementary material. However, despite this strong foundation, local details in the rendered depth maps, such as the dinosaur horn (indicated by the yellow dashed box in Fig. 6), still contain severe holes and artifacts. This limitation necessitates the introduction of further geometric constraints.

Effectiveness of Stereo Depth Regularization. Building upon the dense initialization, the integration of stereo depth regularization further refines the geometry. Table 5 demonstrates consistent quantitative improvements across all

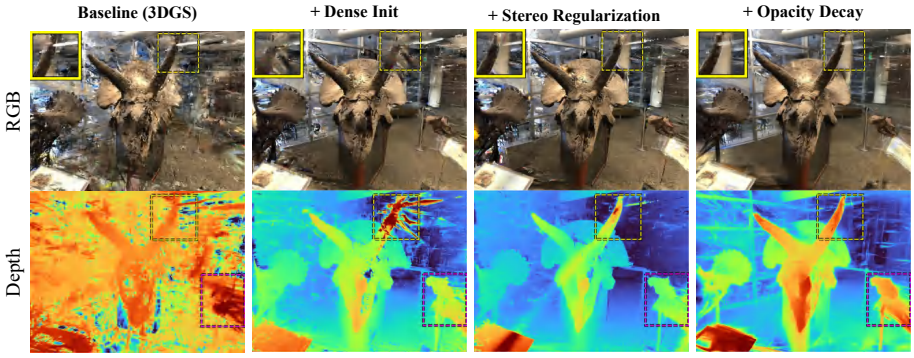

Fig. 6: Ablation Study of Visual Quality. We compare the rendered RGB images and depth maps of novel views to verify the contributions of each proposed component.

metrics. More importantly, Fig. 6 illustrates the substantial qualitative impact of this module: severe depth holes on the horn are effectively filled, and the background geometry becomes significantly smoother. To verify that these improvements arise from the proposed regularization formulation rather than a specific network architecture, we evaluate the results of different stereo matching models in the supplementary material.

Effectiveness of Gradient-Aware Opacity Decay. Finally, the replacement of the standard periodic reset with the proposed gradient-aware opacity decay strategy yields the most significant visual and quantitative refinements. This module achieves a best performance of 21.91 PSNR on the LLFF dataset and 21.46 PSNR on the DTU dataset, and it excels at pruning redundant noise. As shown in the final column of Fig. 6, the remaining floaters in the background and around the boundaries of the foreground are cleanly eliminated. By dynamically evaluating the importance of each Gaussian through the relative gradient, the proposed strategy successfully preserves essential surface structures while pruning noise points. To rigorously justify our opacity decay formulation, we compare our exponential decay function against alternative mapping strategies (*e.g.*, constant, step, and linear decay) and raw gradient inputs in the supplementary material.

5 Conclusion

In this paper, we presented StereoGS, a novel framework that effectively addresses the critical challenges of scale ambiguity and geometric inconsistency in sparse-view 3D Gaussian Splatting. By seamlessly integrating foundation stereo models, we introduced a Stereo Depth Regularization mechanism that constructs virtual stereo pairs to enforce absolute scale and binocular consistency during optimization. Furthermore, we designed a Gradient-Aware Opacity Decay strategy that dynamically prunes redundant primitives based on their relative optimiza-

tion gradients, preserving essential geometry while eliminating noise. Coupled with a Consistency-Aware Dense Initialization, StereoGS establishes a robust geometric foundation from the outset. Extensive experiments across multiple standard benchmarks demonstrate that StereoGS achieves state-of-the-art novel view synthesis quality under sparse-view conditions, all while maintaining the real-time rendering efficiency of vanilla 3DGS.

Limitations. Stereo models may produce inaccurate predictions in extremely textureless regions. Furthermore, online stereo depth regularization inevitably incurs additional training overhead. Addressing these challenges represents a promising direction for future research.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021) 1, 3
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022) 3, 8
3. Cao, C., Ren, X., Fu, Y.: Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. arXiv preprint arXiv:2208.02541 (2022) 8, 20, 29
4. Chung, J., Oh, J., Lee, K.M.: Depth-regularized optimization for 3d gaussian splatting in few-shot images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 811–820 (2024) 2, 4, 7
5. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4
6. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017) 6
7. Han, L., Zhou, J., Liu, Y.S., Han, Z.: Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. Advances in Neural Information Processing Systems **37**, 68595–68621 (2024) 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 22, 25
8. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2), 328–341 (2008) 6
9. Izquierdo, S., Sayed, M., Firman, M., Garcia-Hernando, G., Turmukhambetov, D., Civera, J., Mac Aodha, O., Brostow, G.J., Watson, J.: MVSAnywhere: Zero shot multi-view stereo. In: CVPR (2025) 3, 8, 20, 27, 28
10. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014) 8
11. Jie, Z., Wang, P., Ling, Y., Zhao, B., Wei, Y., Feng, J., Liu, W.: Left-right comparative recurrent model for stereo matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3838–3846 (2018) 6

12. Jing, J., Luo, W., Mao, Y., Mikolajczyk, K.: Lite any stereo: Efficient zero-shot stereo matching. arXiv preprint arXiv:2511.16555 (2025) 20, 22
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 1, 3, 4, 7, 8, 9, 10, 11, 12
14. Kheradmand, S., Rebain, D., Sharma, G., Sun, W., Tseng, Y.C., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: 3d gaussian splatting as markov chain monte carlo. NeurIPS (2024) 4
15. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12912–12921 (2022) 4
16. Lee, J.C., Rho, D., Sun, X., Ko, J.H., Park, E.: Compact 3d gaussian representation for radiance field. In: CVPR (2024) 4
17. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: CVPR (2024) 2, 4, 5, 7, 8, 9, 10, 11, 12
18. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (ToG) **38**(4), 1–14 (2019) 8
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 1, 3, 8
20. Min, J., Jeon, Y., Kim, J., Choi, M.: S2m2: Scalable stereo matching model for reliable depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26729–26739 (2025) 20, 22
21. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**, 102:1–102:15 (2022) 3
22. Niedermayr, S., Stumpfegger, J., Westermann, R.: Compressed 3d gaussian splatting for accelerated novel view synthesis. In: CVPR (2024) 4
23. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022) 1, 4, 8, 9, 10, 11, 12
24. Paliwal, A., Ye, W., Xiong, J., Kotovenko, D., Ranjan, R., Chandra, V., Kalantari, N.K.: Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In: European Conference on Computer Vision. pp. 19–37. Springer (2024) 4
25. Park, H., Ryu, G., Kim, W.: Dropgaussian: Structural regularization for sparse-view gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21600–21609 (2025) 4, 9, 11, 12, 27
26. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 5
27. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020) 5
28. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4

29. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) 7, 20, 29
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 7, 24
31. Song, J., Park, S., An, H., Cho, S., Kwak, M.S., Cho, S., Kim, S.: Därf: Boosting radiance fields from sparse inputs with monocular depth adaptation. In: NIPS (2023) 4
32. Song, M., Lin, X., Zhang, D., Li, H., Li, X., Du, B., Qi, L.: D^2 gs: Depth-and-density guided gaussian splatting for stable and accurate sparse-view reconstruction (2026) 9, 11
33. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022) 3
34. Truong, P., Danelljan, M., Timofte, R., Van Gool, L.: Pdc-net+: Enhanced probabilistic dense correspondence network. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 10247–10266 (2023) 8, 20, 29
35. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 1, 4, 8, 9, 10, 11, 12
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9
37. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5249–5260 (2025) 2, 5, 6, 20, 22, 27
38. Wynn, J., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4
39. Xu, W., Gao, H., Shen, S., Peng, R., Jiao, J., Wang, R.: Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. In: European Conference on Computer Vision. pp. 203–220. Springer (2024) 2, 4, 7, 8, 9, 10, 11
40. Yan, Z., Low, W.F., Chen, Y., Lee, G.H.: Multi-scale 3d gaussian splatting for anti-aliased rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20923–20931 (2024) 4
41. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR) (2023) 3, 4, 8, 9, 10, 11, 12
42. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024) 5
43. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018) 8, 27
44. Ye, Z., Li, W., Liu, S., Qiao, P., Dou, Y.: Absgs: Recovering fine details in 3d gaussian splatting. In: ACM MM (2024) 4
45. Zhang, J., Li, J., Yu, X., Huang, L., Gu, L., Zheng, J., Bai, X.: Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In: European Conference on Computer Vision. pp. 335–352. Springer (2024) 4, 9, 10, 11, 12

46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 9
47. Zhang, Z., Hu, W., Lao, Y., He, T., Zhao, H.: Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. In: ECCV (2024) 4
48. Zheng, Y., Jiang, Z., He, S., Sun, Y., Dong, J., Zhang, H., Du, Y.: Nexusgs: Sparse view synthesis with epipolar depth priors in 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26800–26809 (2025) 2, 7, 9, 10, 11, 12
49. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: CVPR (2023) 4
50. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: European Conference on Computer Vision. Springer (2024) 1, 2, 4, 5, 7, 8, 9, 10, 11, 12